7.91 – Lecture #1 Michael Yaffe

Introduction to Bioinformatics

Focus on Kinases

& Pairwise Sequence Comparisons

```
ARDFSHGLLENKLLGCDSMRWE
 :  :  :  :  :  :  :
GRDYKMALLEQWILGCD-MRWD
```

Reading:
This lecture: Mount pp.1-7, 29-35, 45-48, 51-64
Next lecture: Mount pp. 8-9, 65-89, 96-115, +more

The Protein Data Bank (PDB - http://www.pdb.org/) is the single worldwide repository for the processing and distribution of 3-D biological macromolecular structure data.

Berman, H. M., J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The Protein Data Bank. *Nucleic Acids Research* 28 (2000): 235-242.

(PDB Advisory Notice on using materials available in the archive: http://www.rcsb.org/pdb/advisory.html)

Outline

- Review of biological fundamentals
- Genetics example
- Biological databases/NCBI resources
- Simplified sequence analysis – where to start…
- Simple sequence comparisons
- Definitions of related sequences
- Concepts and types of alignments – the good, the bad, and the ugly
- Dot matrix alignments
- Computational efficiency
- Recursion and dynamic programming
- Substitution matrices: PAM, BLOSUM, Gonnet

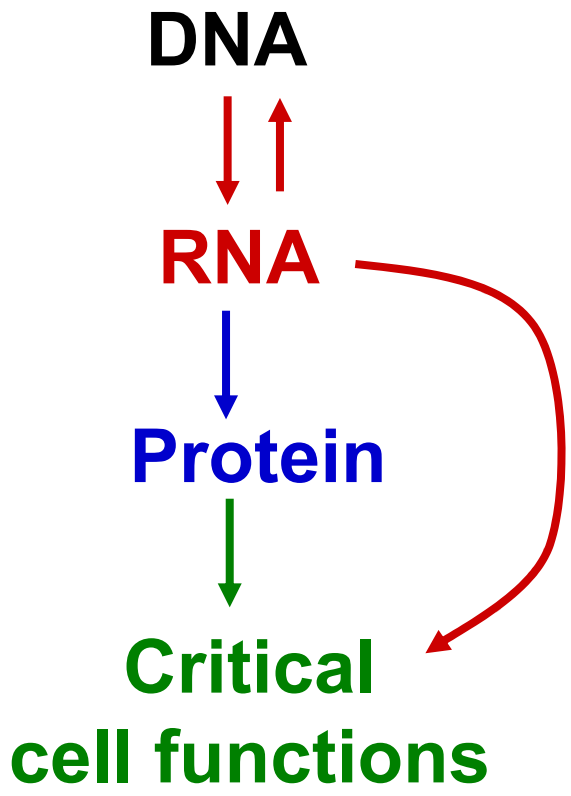
Outline (cont)

- Gaps
- Applied dynamic programming: global alignments: Needleman-Wunsch
- Applied dynamic programming: local alignments – Smith-Waterman
- Basic statistics of sequence alignments

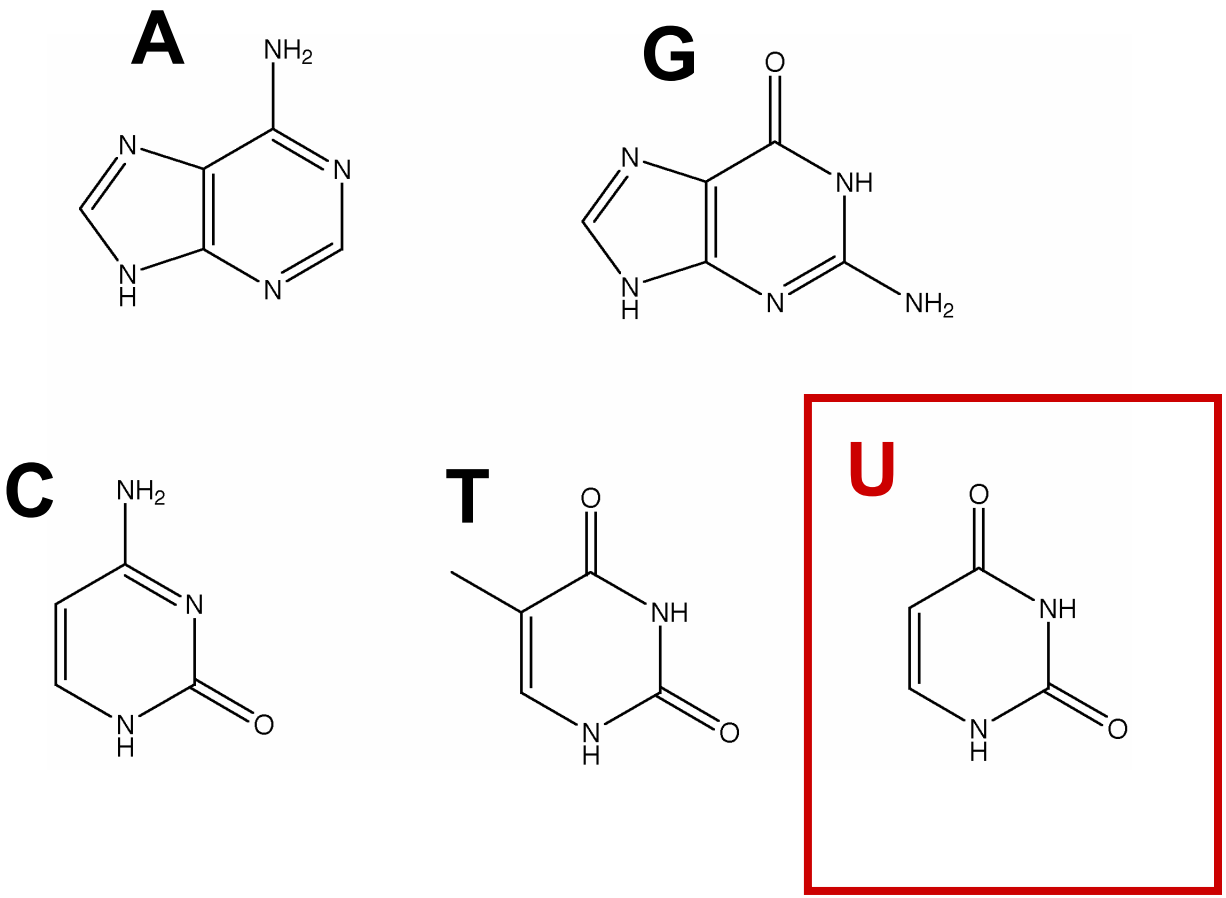
Review of biological fundamentals

Genetic material

- gene as a concept
- DNA as hereditary material

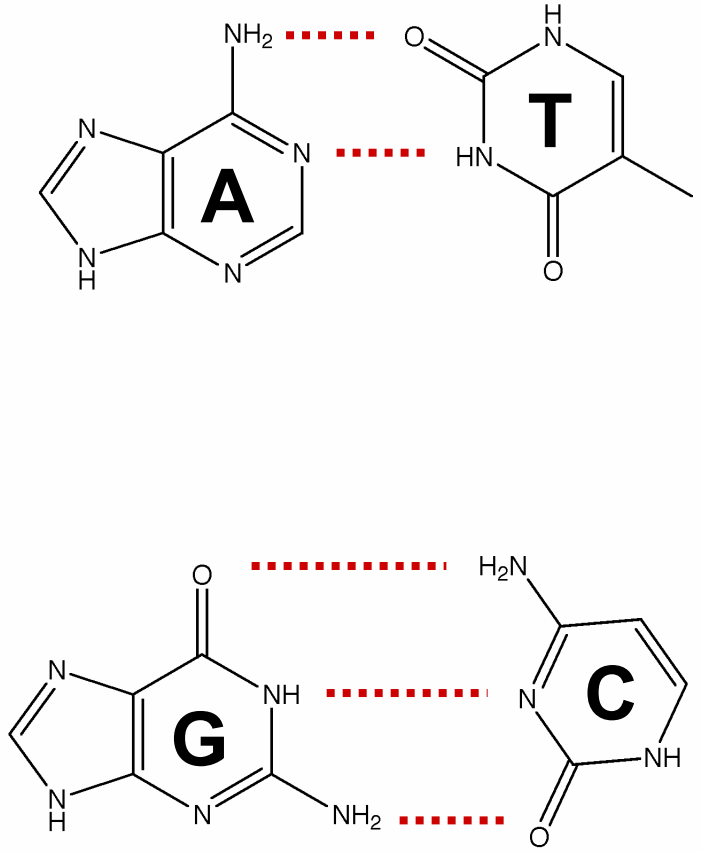


DNA structure



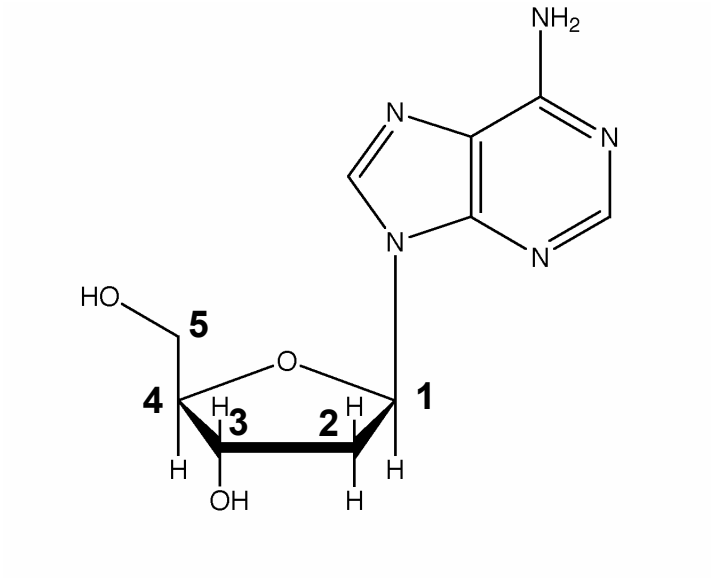
Bases

DNA structure

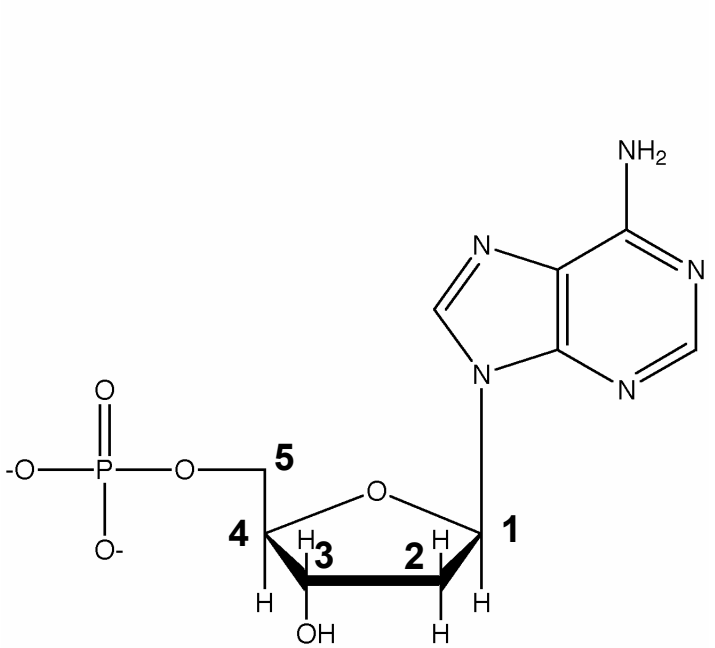


Base pairing

DNA structure

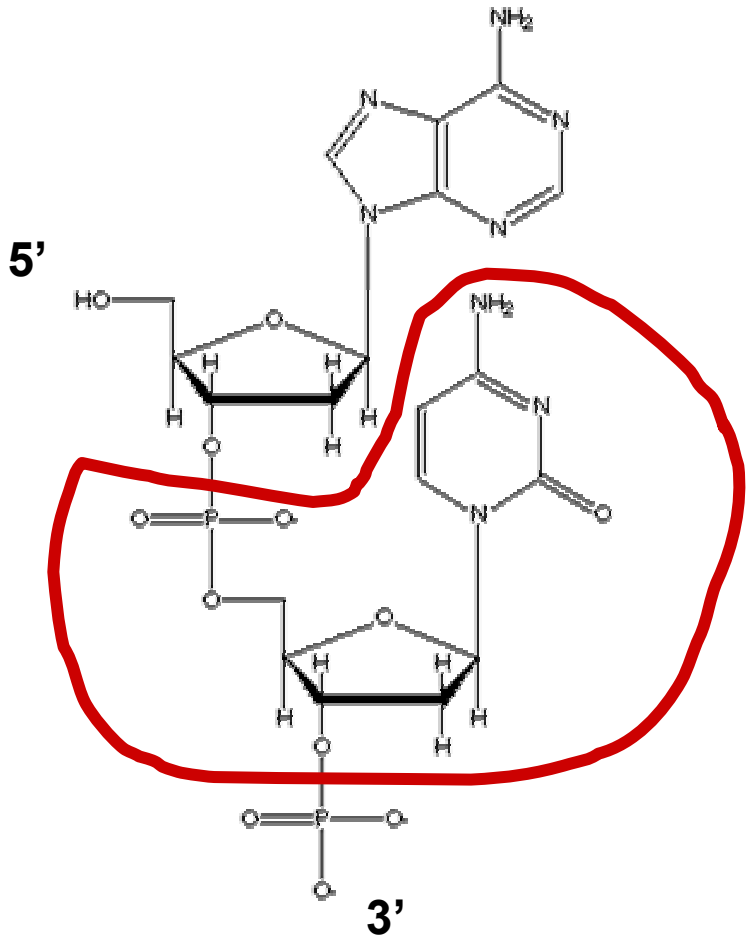


Nucleoside

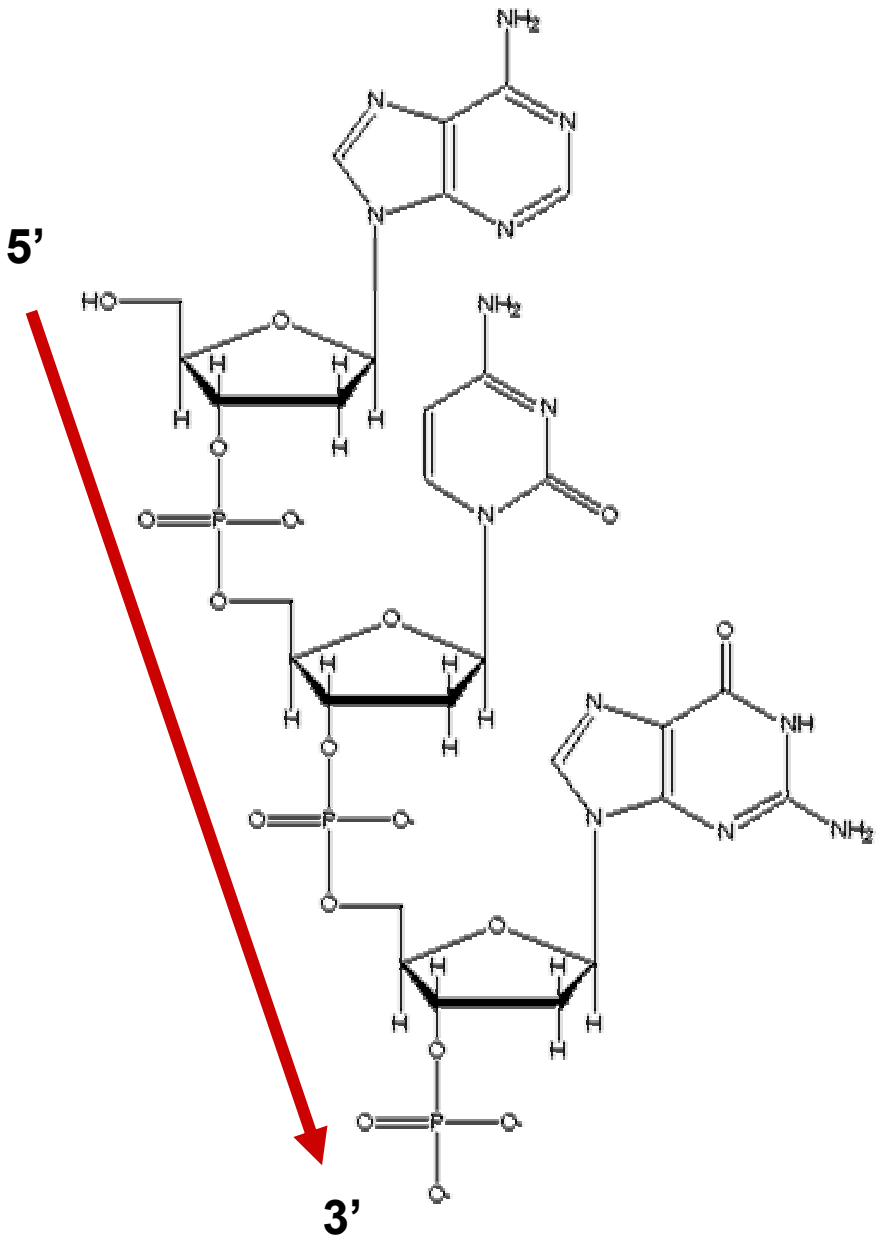


Nucleotide

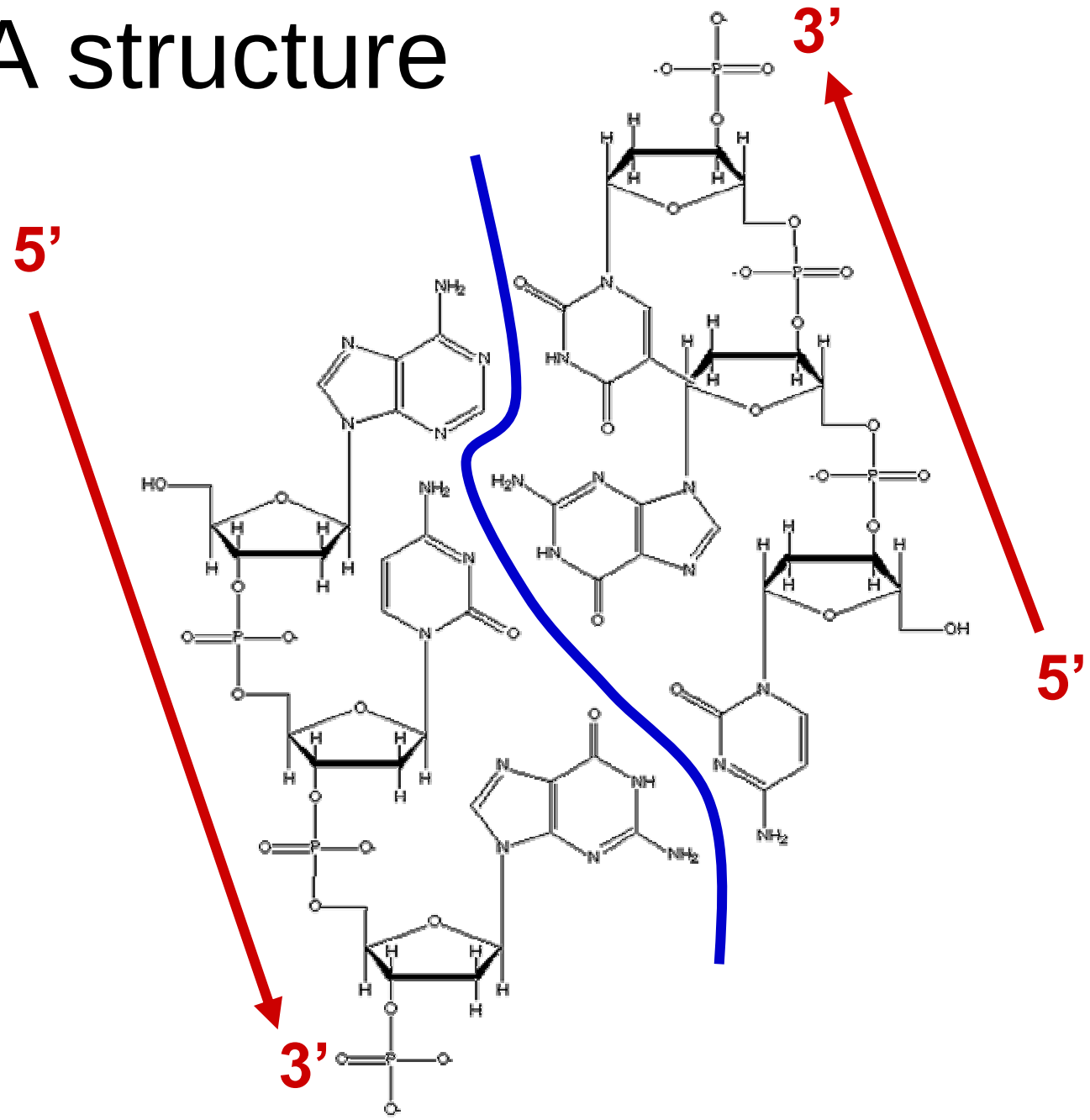
DNA structure



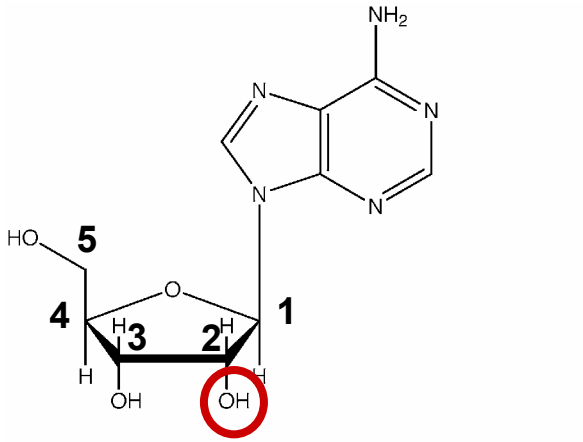
DNA structure

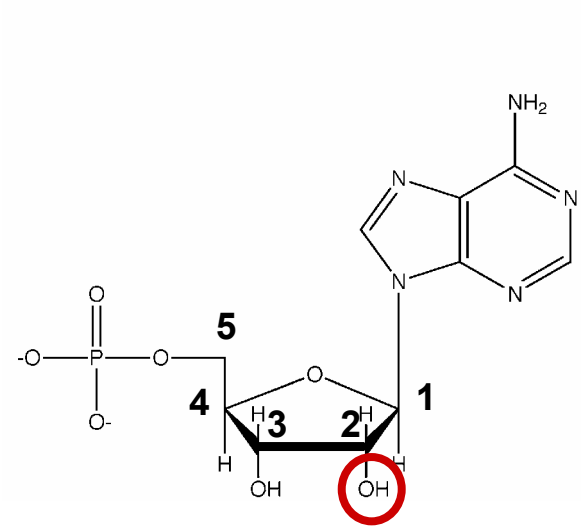


DNA structure

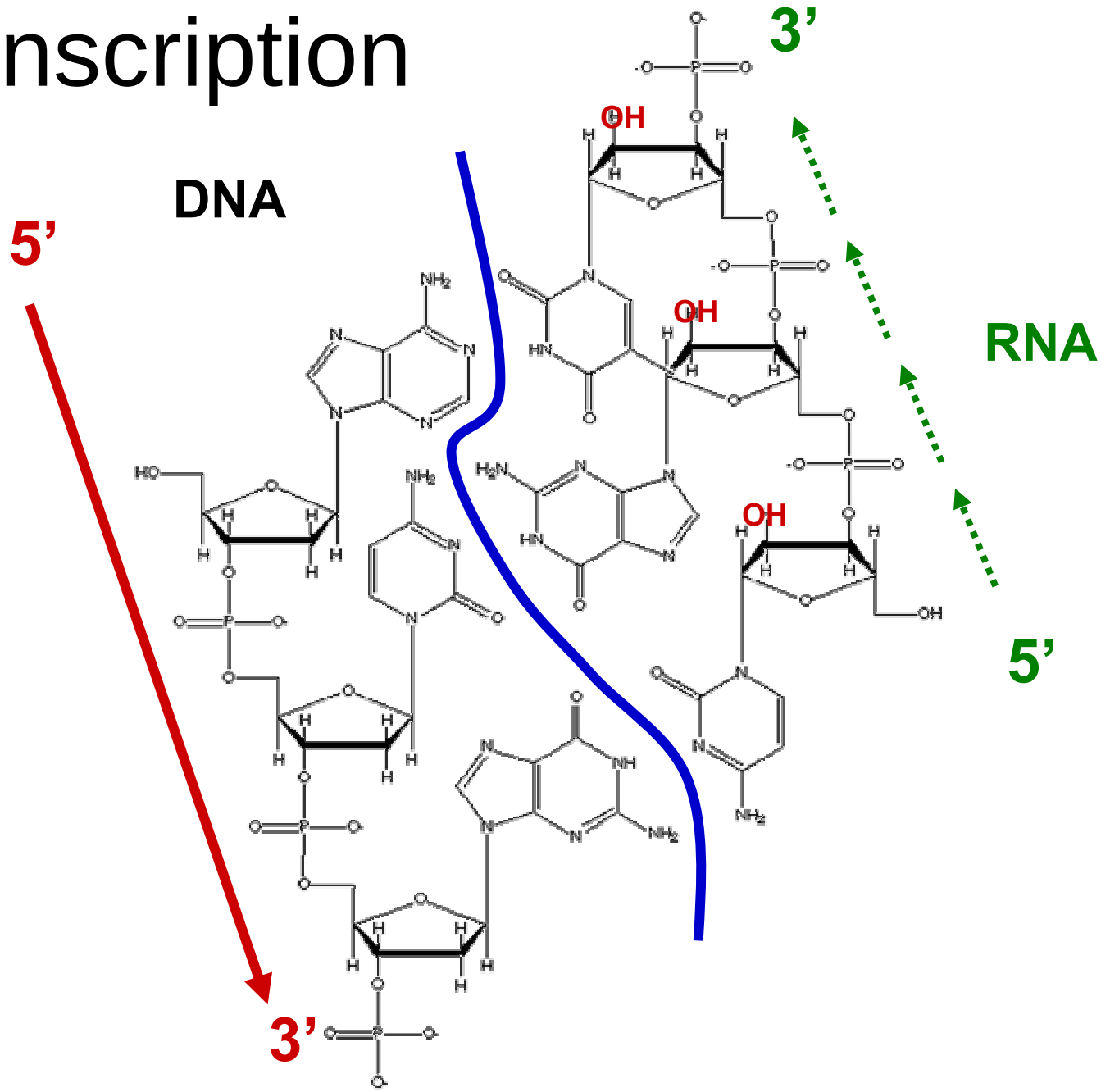


RNA structure

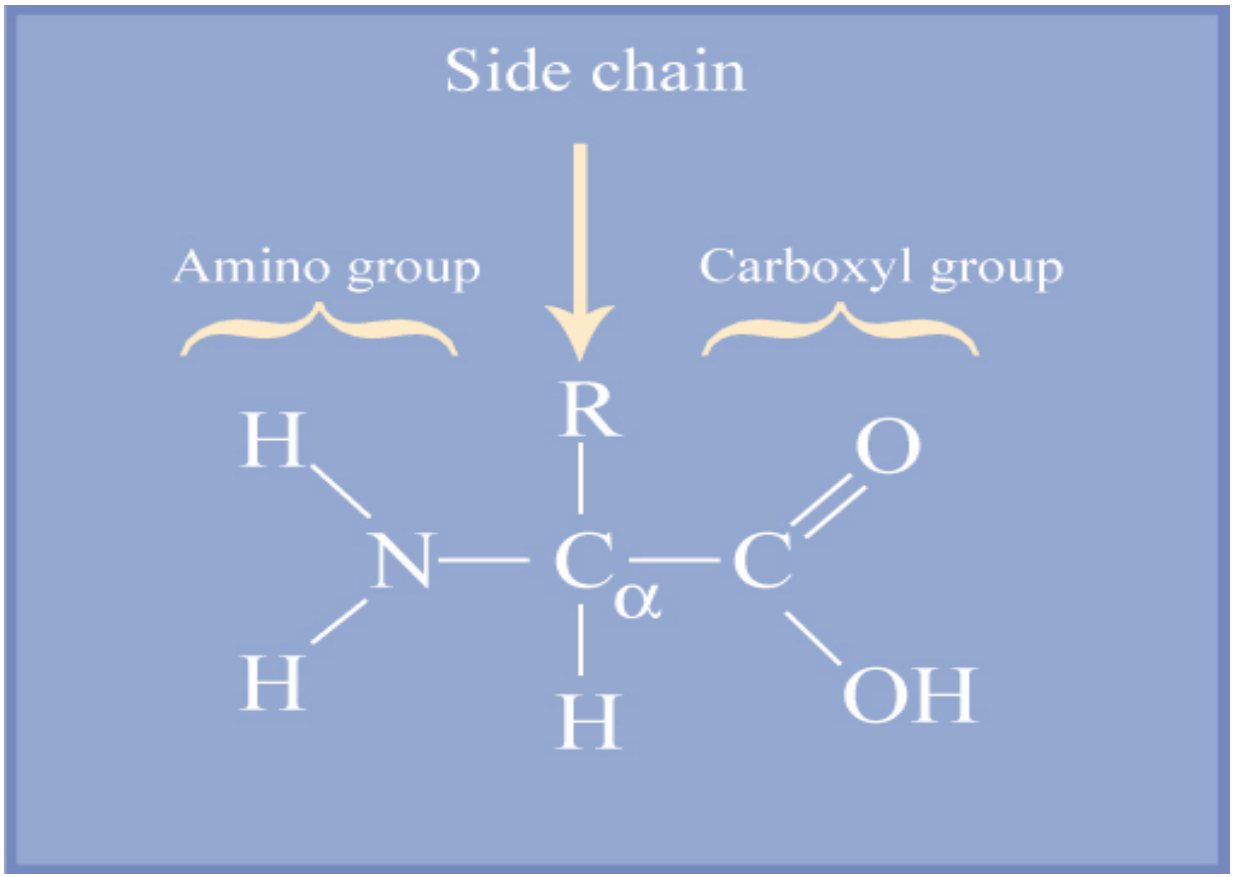




Transcription

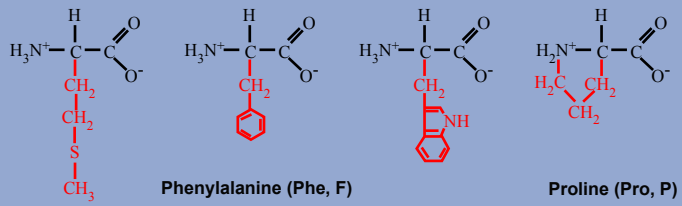


Protein structure



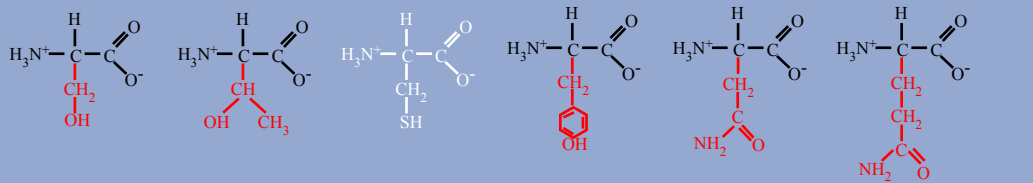
Nonpolar, Hydrophobic R-groups

Glycine (Gly, G) Alanine (Ala, A) Valine (Val, V) Leucine (Leu, L) Isoleucine (Ile, I)



Methionine (Met, M) Phenylalanine (Phe, F) Tryptophan (Trp, W) Proline (Pro, P)

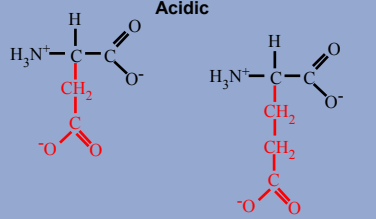
Polar, Hydrophillic R-groups



Serine (Ser, S) Threonine (Thr, T) Cysteine (Cys, C) Tyrosine (Tyr, Y) Asparagine (Asn, N) Glutamine (Gln, Q)

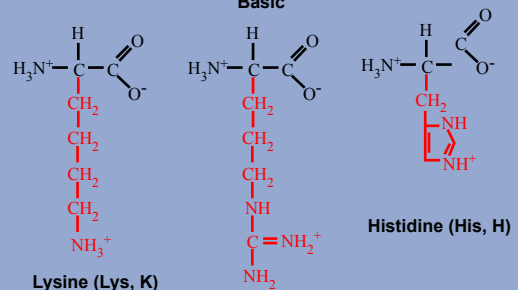
Electrically charged

Acidic



Aspartic acid (Asp, D) Glutamic acid (Glu, E)

Basic



Lysine (Lys, K) Arginine (Arg, R) Histidine (His, H)

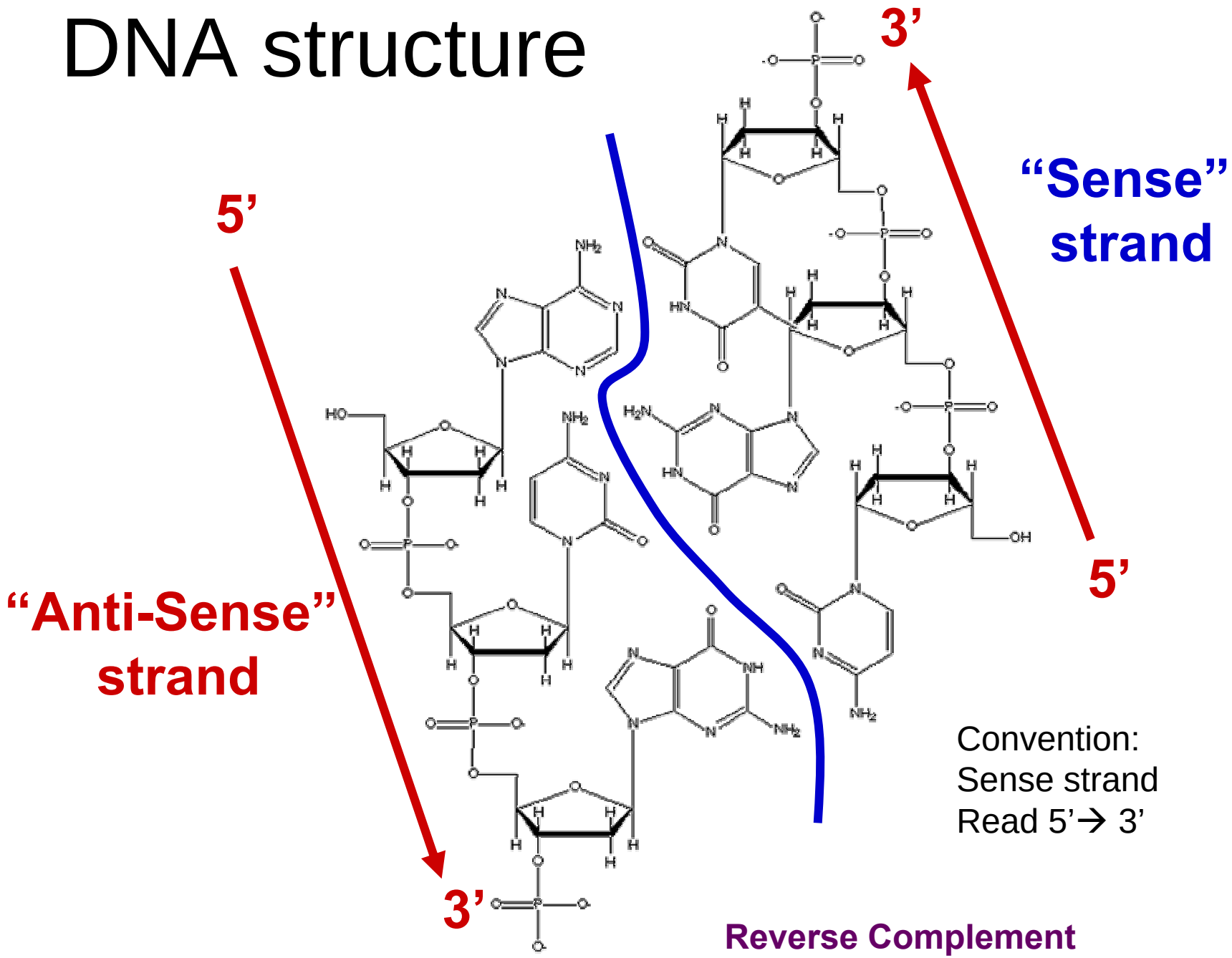
Codon Table

(5')....pNpNpN....(3') in mRNA

Middle Base of Codon →

Base at 5' End of Codon ↓	U	C	A	G	Base at 3' End of Codon ↓
U	phe (UUU)	ser	tyr	cys	U
	phe	ser	tyr	cys	C
	leu	ser	termination	termination	A
	leu	ser	termination	trp	G
C	leu	pro	his	arg	U
	leu	pro	his	arg	C
	leu	pro	gln	arg	A
	leu	pro	gln	arg	G
A	ile	thr	asn	ser	U
	ile	thr	asn	ser	C
	ile	thr	lys	arg	A
	met (and initiation)	thr	lys	arg	G
G	val	ala	asp	gly	U
	val	ala	asp	gly	C
	val	ala	glu	gly	A
	val	ala	glu	gly	G

DNA structure



Genetics experiment:

Isolate a yeat mutant that has increased chromosome number

phenotype

Can rescue the phenotype with a piece of DNA

…corresponds to a site of mutation in the yeasts DNA

genotype

CGTTTTTCCTGTAAAGCGCTTAATTTGTTTACCATTCTATAAAAACCTTGAGCTAAGGCCAACTGATGCA
ATTGCTCAAGTGAATGCATAAACAAAGCAAGATCATTCTTAGCGCAAAAAAAACTGGGATTTTGAAATAC
AACAAAAGAAAGAAGTAAAAAGGGAATGCAACGCAATAGTTTAGTAAATATCAAACTAAACGCTAATTCG
CCATCGAAAAAGACCACAACAAGACCAAATACGTCCAGGATCAATAAACCATGGAGAATATCCCATTCGC
CGCAGCAAAGAAACCCGAATTCAAAAATACCTTCACCTGTAAGAGAAAAATTGAACAGATTACCTGTAAA
CAATAAGAAGTTTTTGGATATGGAAAGCTCCAAAATTCCATCACCTATAAGGAAAGCGACTTCTTCCAAA
ATGATACACGAAAATAAGAAGCTACCTAAATTTAAATCCCTATCACTCGATGACTTTGAACTGGGGAAGA
AATTAGGAAAGGGTAAATTCGGTAAAGTTTATTGCGTTCGGCACAGGAGTACAGGATATATTTGCGCACT
GAAAGTAATGGAGAAGGAAGAAATAATAAAGTATAATTTACAGAAACAATTCAGAAGGGAGGTAGAAATA
CAAACATCGCTAAATCATCCGAATCTAACTAAATCATACGGCTATTTTCATGATGAAAAAAGAGTGTACC
TGCTAATGGAATACTTAGTCAATGGGGAAATGTATAAACTATTGAGGTTACACGGACCCTTCAACGATAT
TTTAGCATCAGATTATATTTATCAAATTGCCAATGCCCTAGATTATATGCATAAAAAGAATATTATTCAT
AGAGATATTAAACCTGAAAATATACTAATAGGGTTCAATAATGTCATTAAGTTAACGGACTTCGGATGGA
GTATAATAAATCCGCCAGAAAATAGAAGGAAAACTGTCTGTGGGACAATTGACTACCTTTCTCCAGAAAT
GGTGGAGTCAAGGGAATATGATCACACTATAGATGCATGGGCTCTTGGCGTCCTGGCGTTTGAACTACTG
ACCGGTGCCCCTCCGTTCGAAGAAGAAATGAAAGATACTACATATAAAAGGATAGCAGCACTGGATATCA
AAATGCCCAGTAACATTTCTCAGGATGCGCAAGATTTAATACTTAAACTACTAAAATACGACCCCAAAGA
TAGAATGCGCCTTGGAGACGTAAAAATGCATCCTTGGATACTAAGAAACAAGCCCTTTTGGGAAAATAAG
CGGTTATAGAATTAAAGTATGACAGAATCGTTTGAAGGGCACTATTAATCACTCCCGCACATATCACATA
ATACTAAGTATCCATTTCTAATATTTCATTACTCTTTTCGGCATCGTATATTGCGATATTTGATTAAATT
TTCTTGTTCATTTTTTCCTCTTTTCTTTCGCCTTTGTCGAAAGAAAAGAGGAAAACAAGCTGAAAATTGC
TATGCATTAAAGTAGCAGATTTACTTTGTTGAGTTGGTTCTGATCAATAATAAGAGTAATGAAAGAAAGC
AAAAAAATGGCTAAAGATAATTTAACTAATTTGCTCTCTCAATTGAACATTCAATTGTCTCAA

The National Center for Biotechnology Information

- Created as a part of NLM in 1988
  - Establish public databases
  - Perform research in computational biology
  - Develop software tools for sequence analysis
  - Disseminate biomedical information

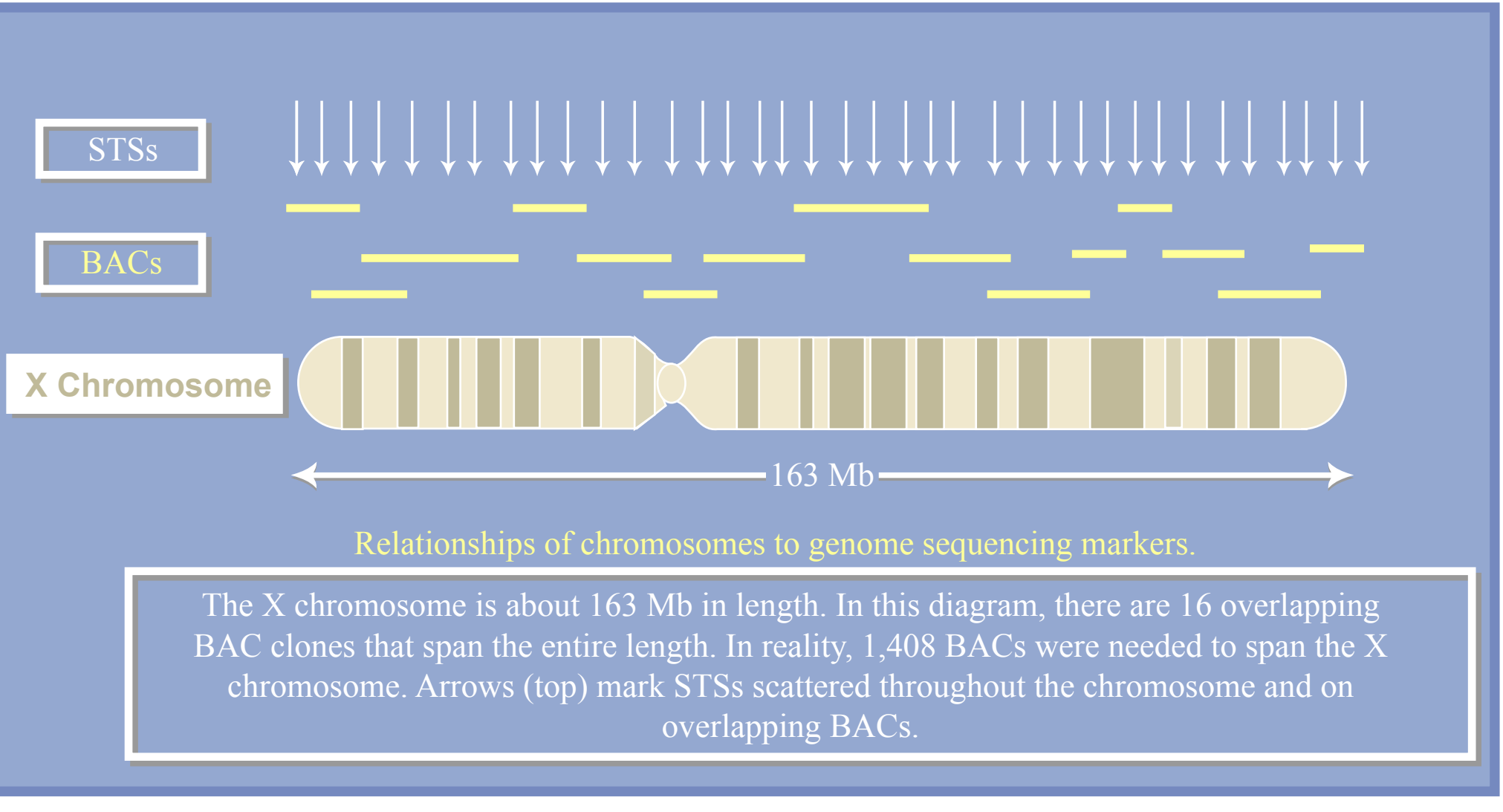
Molecular Databases

- Primary Databases
  - Original submissions by experimentalists
  - Database staff organize but don't add additional information
    - **Example: GenBank**
- Derivative Databases
  - Human curated
    - compilation and correction of data
    - **Example: SWISS-PROT, NCBI RefSeq mRNA**
  - Computationally Derived
    - **Example: UniGene**
  - Combinations
    - **Example: NCBI Genome Assembly**

What is GenBank?

NCBI's Primary Sequence Database

- **Nucleotide only sequence database**
- **Archival in nature**
- **GenBank Data**
  - **Direct submissions individual records (BankIt, Sequin)**
  - **Batch submissions via email (EST, GSS, STS)**
  - **ftp accounts sequencing centers**
- **Data shared three collaborating databases**
  - **GenBank**
  - **DNA Database of Japan (DDBJ).**
  - **European Molecular Biology Laboratory Database (EMBL) at EBI.**



Relationships of chromosomes to genome sequencing markers.

The X chromosome is about 163 Mb in length. In this diagram, there are 16 overlapping BAC clones that span the entire length. In reality, 1,408 BACs were needed to span the X chromosome. Arrows (top) mark STSs scattered throughout the chromosome and on overlapping BACs.

GenBank: NCBI's Primary Sequence Database

Release 133	December 2003
19,808,101	Records
28,507,990,166	Nucleotides
110,000 +	Species

- full release every two months
- incremental and cumulative updates daily
- available only through internet

ftp://ftp.ncbi.nih.gov/genbank/

107.07 Gigabytes of data

GenBank Divisions

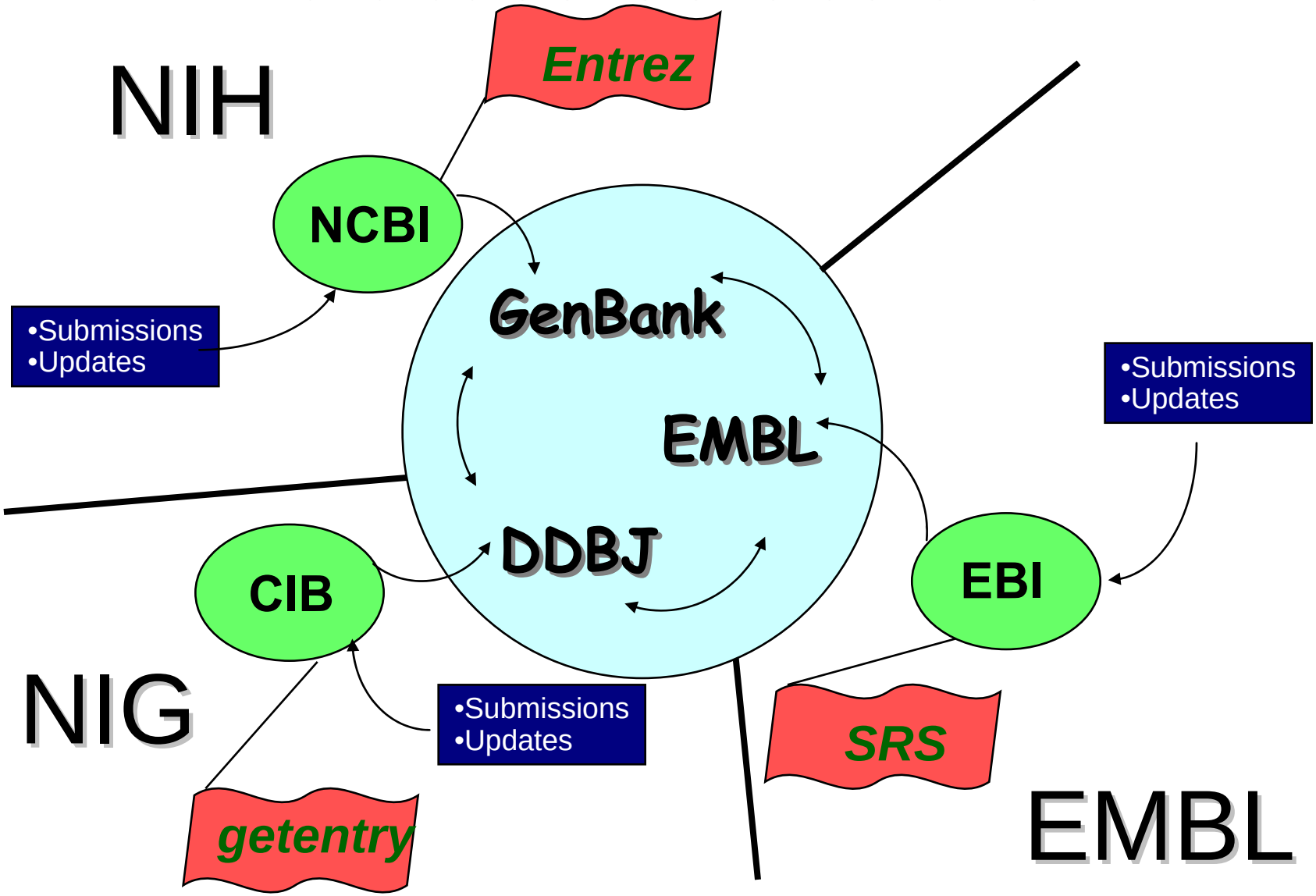
Bulk Sequence Divisions

PAT	Patent
EST	Expressed Sequence Tags
STS	Sequence Tagged Sites
GSS	Genome Survey Sequences
HTG	High Throughput Genome
HTC	High Throughput cDNA
CON	Contig

Traditional Divisions

BCT INV MAM PHG PLN PRI
ROD SYN UNA VRL VRT

The International Sequence Database Collaboration



Web Access

Text

Entrez

Sequence

BLAST

Structure

VAST

Genetics experiment:

Isolate a yeat mutant that has increased chromosome number

phenotype

Can rescue the phenotype with a piece of DNA

…corresponds to a site of mutation in the yeasts DNA

genotype

CGTTTTTCCTGTAAAGCGCTTAATTTGTTTACCATTCTATAAAAACCTTGAGCTAAGGCCAACTGATGCA
ATTGCTCAAGTGAATGCATAAACAAAGCAAGATCATTCTTAGCGCAAAAAAAACTGGGATTTTGAAATAC
AACAAAAGAAAGAAGTAAAAAGGGAATGCAACGCAATAGTTTAGTAAATATCAAACTAAACGCTAATTCG
CCATCGAAAAAGACCACAACAAGACCAAATACGTCCAGGATCAATAAACCATGGAGAATATCCCATTCGC
CGCAGCAAAGAAACCCGAATTCAAAAATACCTTCACCTGTAAGAGAAAAATTGAACAGATTACCTGTAAA
CAATAAGAAGTTTTTGGATATGGAAAGCTCCAAAATTCCATCACCTATAAGGAAAGCGACTTCTTCCAAA
ATGATACACGAAAATAAGAAGCTACCTAAATTTAAATCCCTATCACTCGATGACTTTGAACTGGGGAAGA
AATTAGGAAAGGGTAAATTCGGTAAAGTTTATTGCGTTCGGCACAGGAGTACAGGATATATTTGCGCACT
GAAAGTAATGGAGAAGGAAGAAATAATAAAGTATAATTTACAGAAACAATTCAGAAGGGAGGTAGAAATA
CAAACATCGCTAAATCATCCGAATCTAACTAAATCATACGGCTATTTTCATGATGAAAAAAGAGTGTACC
TGCTAATGGAATACTTAGTCAATGGGGAAATGTATAAACTATTGAGGTTACACGGACCCTTCAACGATAT
TTTAGCATCAGATTATATTTATCAAATTGCCAATGCCCTAGATTATATGCATAAAAAGAATATTATTCAT
AGAGATATTAAACCTGAAAATATACTAATAGGGTTCAATAATGTCATTAAGTTAACGGACTTCGGATGGA
GTATAATAAATCCGCCAGAAAATAGAAGGAAAACTGTCTGTGGGACAATTGACTACCTTTCTCCAGAAAT
GGTGGAGTCAAGGGAATATGATCACACTATAGATGCATGGGCTCTTGGCGTCCTGGCGTTTGAACTACTG
ACCGGTGCCCCTCCGTTCGAAGAAGAAATGAAAGATACTACATATAAAAGGATAGCAGCACTGGATATCA
AAATGCCCAGTAACATTTCTCAGGATGCGCAAGATTTAATACTTAAACTACTAAAATACGACCCCAAAGA
TAGAATGCGCCTTGGAGACGTAAAAATGCATCCTTGGATACTAAGAAACAAGCCCTTTTGGGAAAATAAG
CGGTTATAGAATTAAAGTATGACAGAATCGTTTGAAGGGCACTATTAATCACTCCCGCACATATCACATA
ATACTAAGTATCCATTTCTAATATTTCATTACTCTTTTCGGCATCGTATATTGCGATATTTGATTAAATT
TTCTTGTTCATTTTTTCCTCTTTTCTTTCGCCTTTGTCGAAAGAAAAGAGGAAAACAAGCTGAAAATTGC
TATGCATTAAAGTAGCAGATTTACTTTGTTGAGTTGGTTCTGATCAATAATAAGAGTAATGAAAGAAAGC
AAAAAAATGGCTAAAGATAATTTAACTAATTTGCTCTCTCAATTGAACATTCAATTGTCTCAA

NCBI

BLAST

PubMed | Entrez | BLAST | OMIM | Taxonomy | Structure

NEW 2 February 2004 BLAST 2.2.7 has been released. Read more...

Info

- FAQs
- News
- References
- Credits

Education

- Program selection guide
- Tutorial
- URL API guide

Download

- Executables
- Databases
- Source code

Support

- Helpdesk
- Mailing list

Nucleotide

- Discontiguous megablast
- Megablast
- Nucleotide-nucleotide BLAST (blastn)
- Search for short, nearly exact matches
- Search trace archives with megablast or discontiguous megablast

Protein

- Protein-protein BLAST (blastp)
- PHI- and PSI-BLAST
- Search for short, nearly exact matches
- Search the conserved domain database (rpsblast)
- Search by domain architecture (cdart)

Translated

- Translated query vs. protein database (blastx)
- Protein query vs. translated database (tblastn)
- Translated query vs. translated database (tblastx)

Genomes

- Human, mouse, rat
- Fugu rubripes, zebrafish
- Insects, nematodes, plants, fungi, malaria
- Microbial genomes, other eukaryotic genomes

Special

- Align two sequences (bl2seq)
- Screen for vector contamination (VecScreen)
- Immunoglobin BLAST (IgBlast)

Meta

- Retrieve results by RID
- Get this page with javascript-free links

Search

AGGGTAAATTCGGTAAAGTTTATTGCGTTCGGCACAGGAGTACACGATATATTTGCGCACT
GGAGAACGAAGAAATAATAAAGTATAATTTACAGAAACAATTCAGAAGGCAGGTAGAAATA
CTAAATCATCCGAATCTAACTAAATCATACGCCTATTTTCATGATGAAAAAAGAGTGTACC
AATACTTACTCAATGGGGAAATGTATAAACTATTCAGGTTACACGGACCCTTCAACGATAT
AGATTATATTTATCAAATTGCCAATGCCCTAGATTATATGCATAAAAAGAATATTATTCAT

Set subsequence From: To:

Choose database nr

Now: BLAST! or Reset query Reset all

Options for advanced blasting

Limit by entrez query or select from: All organisms

Choose filter Low complexity Human repeats Mask for lookup table only Mask lower case

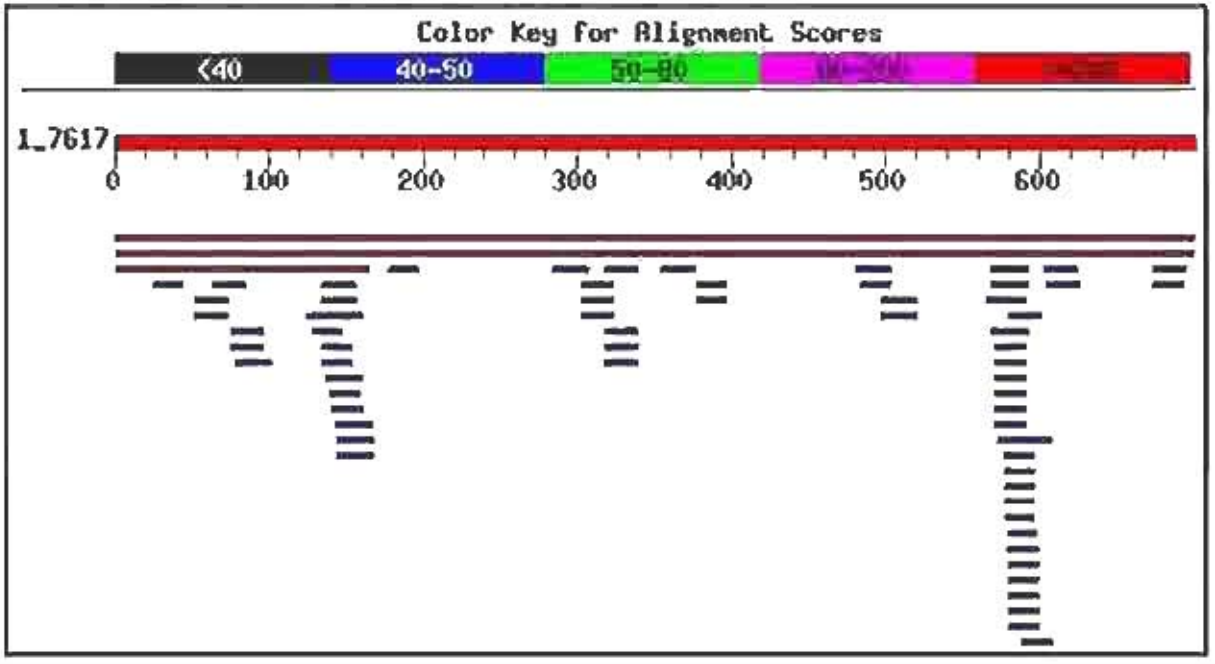
Expect 10

Word Size 11

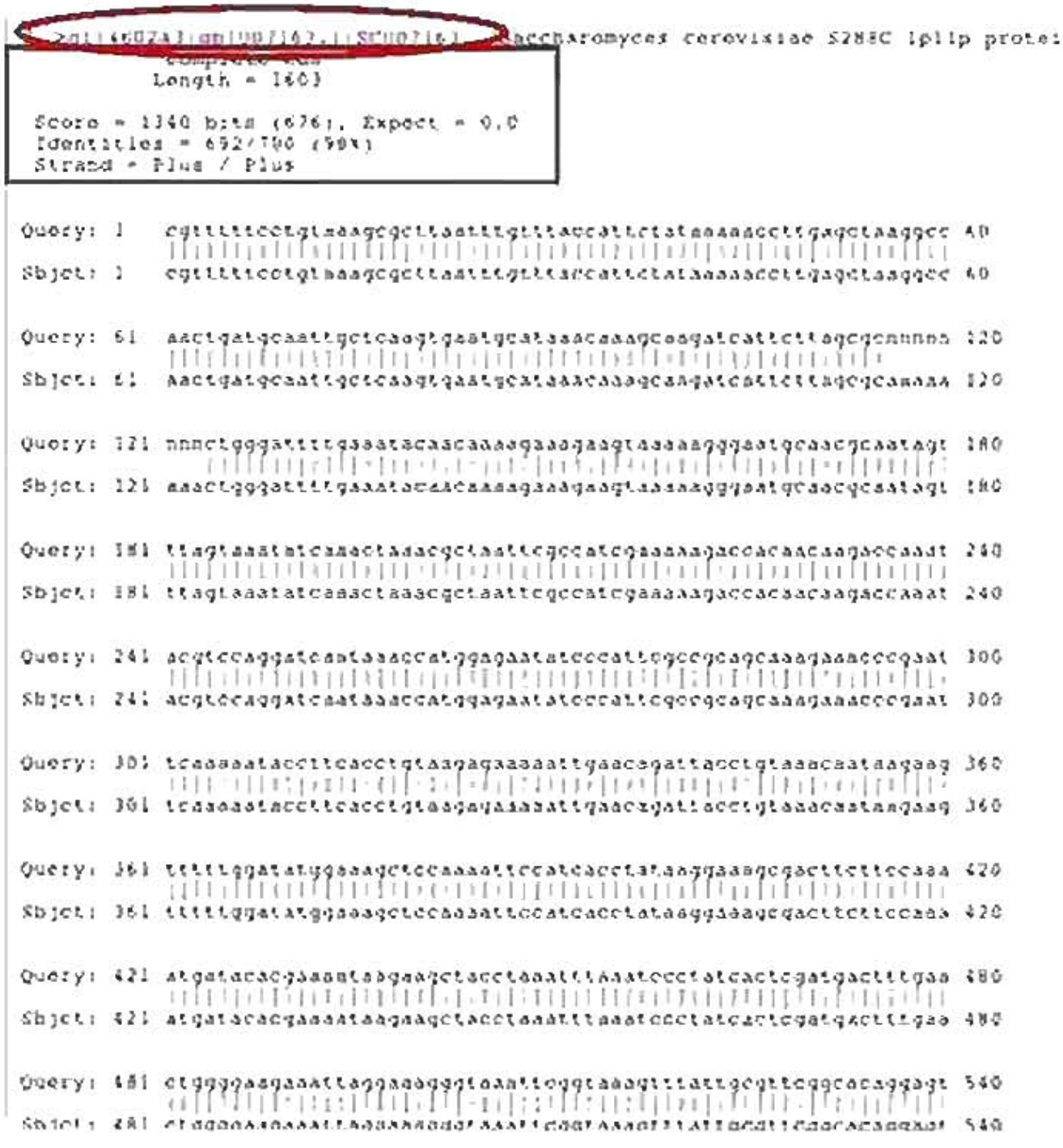
Other advanced

Distribution of 67 Blast Hits on the Query Sequence

Mouse-over to show defline and scores. Click to show alignments



Sequences producing significant alignments:	Score (bits)	E Value
gi|1370433|emb|Z73565.1|SCYPL209C S.cerevisiae chromosome X...	1040	0.0
gi|460243|gb|U07163.1|SCU07163 Saccharomyces cerevisiae S28...	1040	0.0
gi|1370431|emb|Z73564.1|SCYPL208W S.cerevisiae chromosome X...	280	3e-72
gi|14164523|dbj|AP003143.2| Oryza sativa (japonica cultivar...	46	0.089
gi|10257386|dbj|AP002863.1| Oryza sativa (japonica cultivar...	46	0.089
gi|22296697|gb|AC091043.12| Homo sapiens chromosome 18, clo...	46	0.089
gi|27356774|gb|AC115119.5| Mus musculus BAC clone RP23-87M1...	44	0.35
gi|3900845|gb|AC005078.1|AC005078 Homo sapiens BAC clone CT...	44	0.35
gi|37620357|gb|AC146038.2| Pan troglodytes chromosome 7 clo...	44	0.35
gi|24158583|gb|AC121582.3| Mus musculus BAC clone RP23-257I...	42	1.4
gi|38524392|emb|BX004673.10| Zebrafish DNA sequence from cl...	42	1.4


Length = 1603
Score = 1340 bits (676), Expect = 0.0
Identities = 692/700 (99%)
Strand = Plus / Plus
Query: 1 cgttttccctgtaaagcgcttaattgtttaccattctataaaaaccttgagctaaggcc 60
Sbjct: 1 cgttttccctgtaaagcgcttaattgtttaccattctataaaaaccttgagctaaggcc 60
Query: 61 aactgatgcaaattgctcaagtgaatgcataaacaaagcaagatcattcttagcgcnnnnn 120
Sbjct: 61 aactgatgcaaattgctcaagtgaatgcataaacaaagcaagatcattcttagcgcaaaaa 120
Query: 121 nnnctgggatttttgaaatacaacaaaagaaagaagtaaaaagggaatgcaacgcaatagt 180
Sbjct: 121 aaactgggatttttgaaatacaacaaaagaaagaagtaaaaagggaatgcaacgcaatagt 180
Query: 181 ttagtaaatatcaaactaaacgctaattcgccatcgaaaaagaccacaacaagaccaaat 240
Sbjct: 181 ttagtaaatatcaaactaaacgctaattcgccatcgaaaaagaccacaacaagaccaaat 240
Query: 241 acgtccaggatcaataaaccatggagaatatcccattcgccgcagcaaagaaacccgaat 300
Sbjct: 241 acgtccaggatcaataaaccatggagaatatcccattcgccgcagcaaagaaacccgaat 300
Query: 301 tcaaaaataccttcacctgtaagagaaaaattgaacgattacctgtaaacaataagaag 360
Sbjct: 301 tcaaaaataccttcacctgtaagagaaaaattgaacgattacctgtaaacaataagaag 360
Query: 361 ttttggatatggaaagctccaaaattccatcacctataaggaaagcgacttcttccaaa 420
Sbjct: 361 ttttggatatggaaagctccaaaattccatcacctataaggaaagcgacttcttccaaa 420
Query: 421 atgatacacgaaaataagaagctacctaaatttaaatccctatcactcgatgacttgaa 480
Sbjct: 421 atgatacacgaaaataagaagctacctaaatttaaatccctatcactcgatgacttgaa 480
Query: 481 ctgggaagaaattaggaaagggtaaattcggtaaagtttattgcgttcggcacaggagt 540
Sbjct: 481 ctgggaagaaattaggaaagggtaaattcggtaaagtttattgcgttcggcacaggagt 540

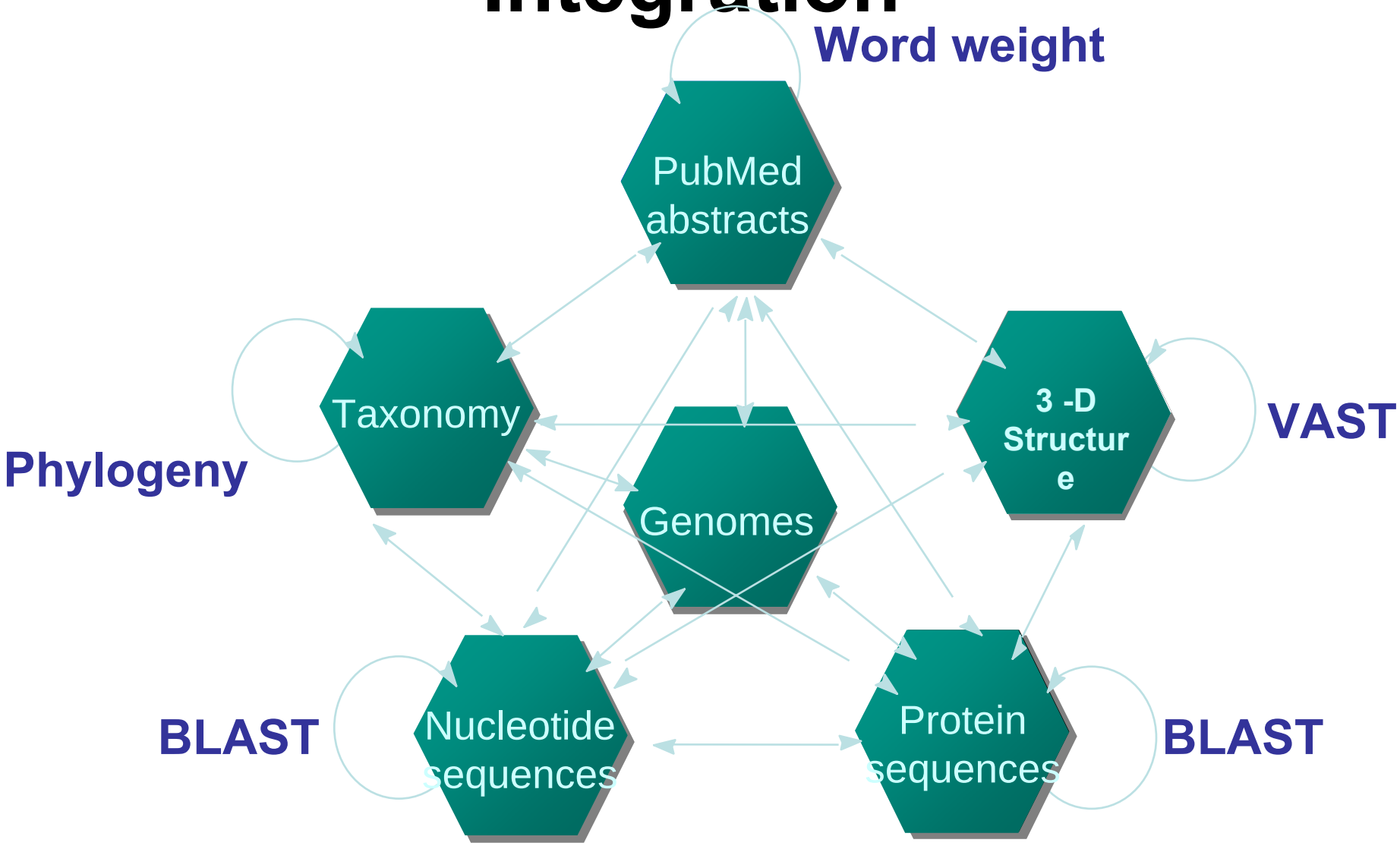
Get selected sequences | Select all | Deselect all

```
  Database: All GenBank+EMBL+DDBJ+PDB sequences (but no EST, STS, GSS,
  or phase 0, 1 or 2 HTGS sequences)
    Posted date:  Jan 28, 2004  9:53 PM
  Number of letters in database: -25,056,781
  Number of sequences in database:  1,740,955

Lambda     K      H
    1.37    0.711     1.31

Gapped
Lambda     K      H
    1.37    0.711     1.31

Matrix: blastn matrix:1 -1
Gap Penalties: Existence: 5, Extension: 2
Number of hits to DB: 242,520
Number of Sequences: 2032434
Number of extensions: 242520
Number of successful extensions: 5340
Number of sequences better than 10.0: 3
Number of HSP's better than 10.0 without gapping: 3
Number of HSP's successfully gapped in prelim test: 0
Number of HSP's that attempted gapping in prelim test: 5330
Number of HSP's gapped (non-prelim): 10
length of query: 1402
length of database: 9,836,170,500
effective HSP length: 22
effective length of query: 678
effective length of database: 9,791,456,952
effective search space: 6638607813456
effective search space used: 6638607813456
T: 0
A: 0
X1: 6 (11.9 bits)
X2: 15 (29.7 bits)
S1: 12 (24.3 bits)
S2: 20 (40.1 bits)
```

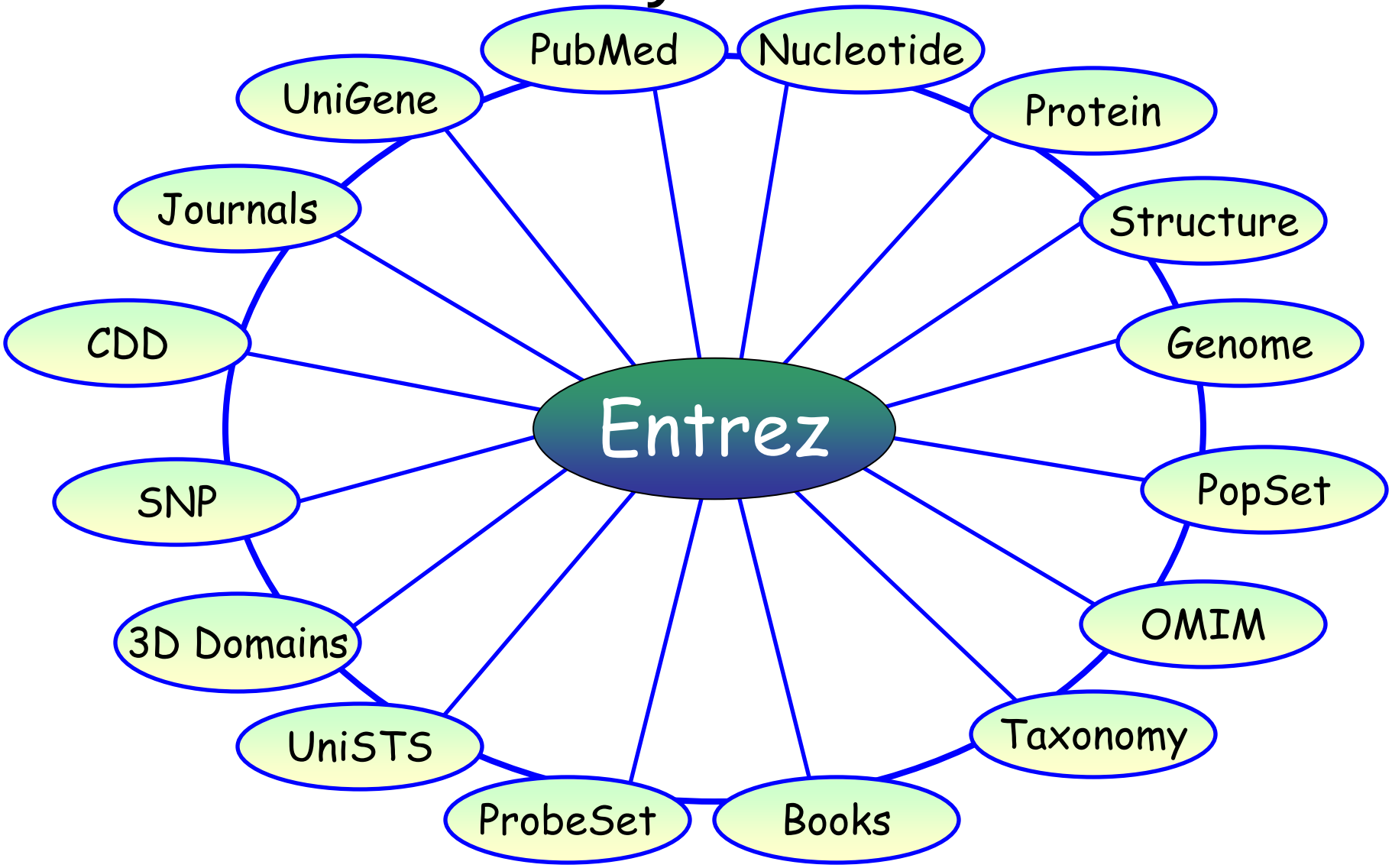
Using Entrez

An integrated database search and retrieval system

Entrez: Database Integration



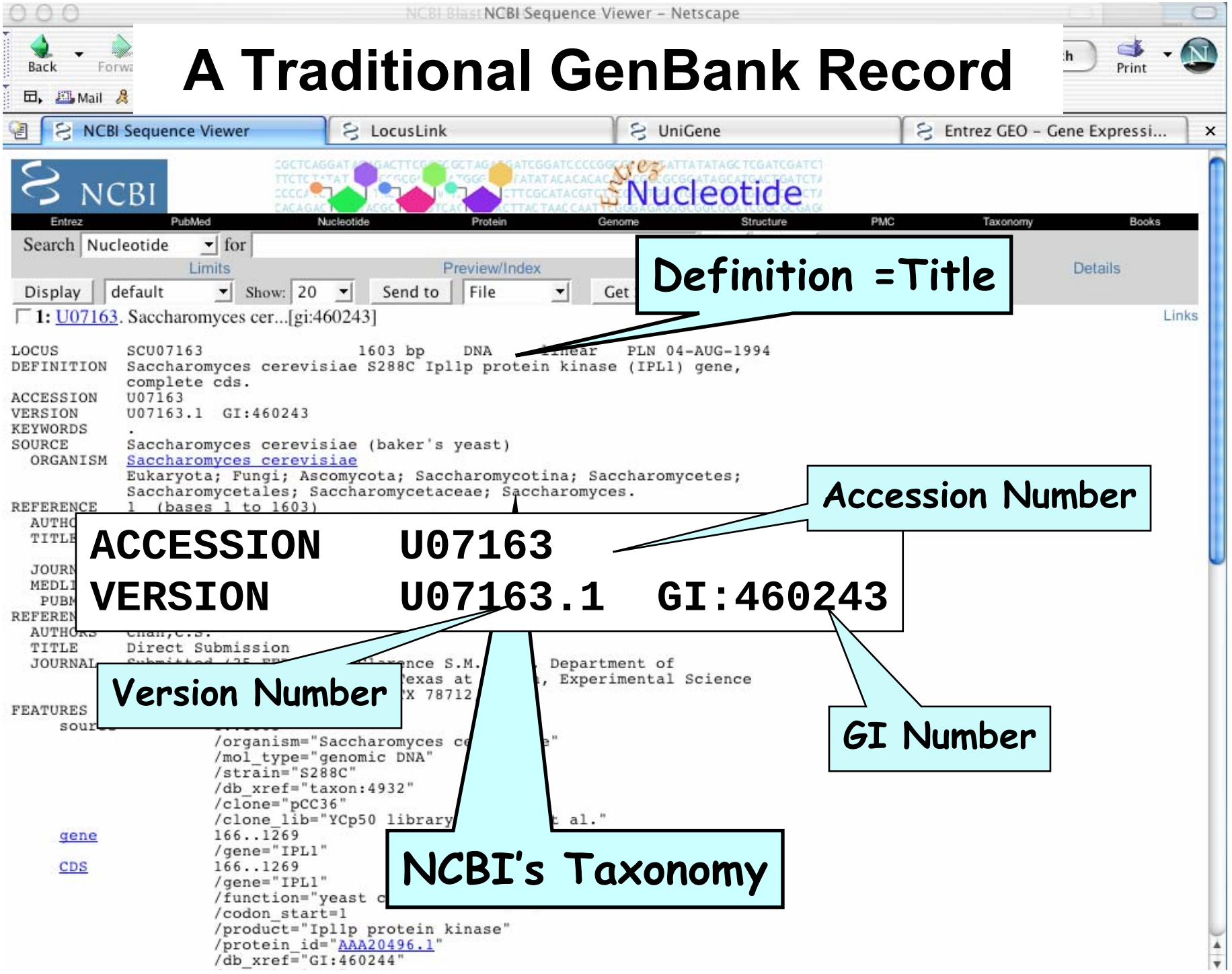
The (ever) Expanding Entrez System



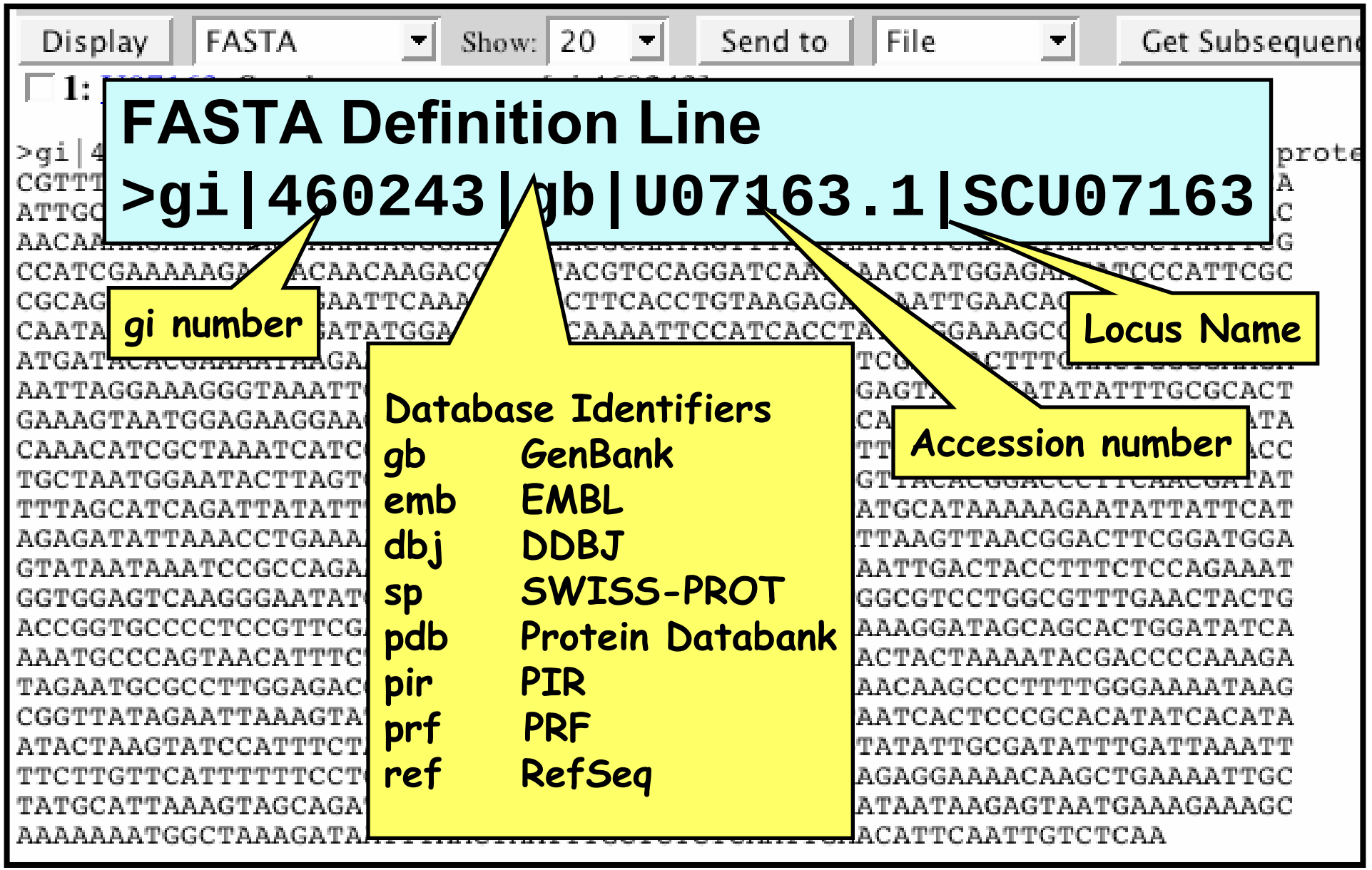
Entrez Databases

PubMed	Biomedical literature
Books	Online textbooks
Nucleotide	GenBank, EMBL, DDBJ, RefSeq, PDB
Protein	[GenBank, EMBL, DDBJ], RefSeq, SWISS-PROT, PIR, PRF, PDB
Genome	Complete genomes
Taxonomy	Organisms in NCBI sequence databases
Structure	MMDB: experimental 3D structures
Domains	CDD: conserved protein domains
3D Domains	Compact 3D protein domains in MMDB
OMIM	Online Mendelian Inheritance in Man
SNP	Single nucleotide polymorphisms
UniSTS	Sequence Tagged Site markers
ProbeSet	Gene expression and microarray datasets
PopSet	Population study datasets
UniGene	Gene-based expressed sequence clusters

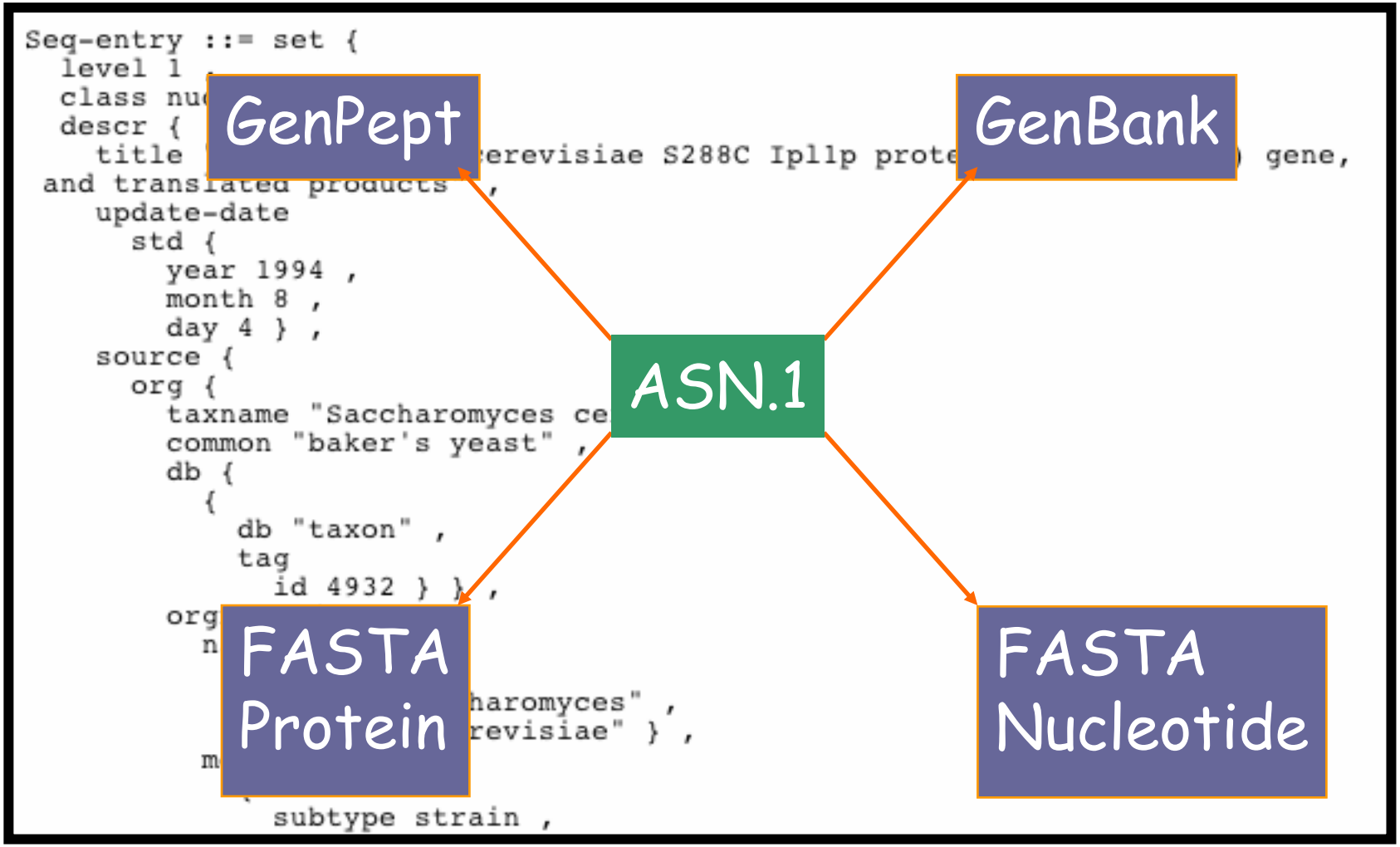
A Traditional GenBank Record



FASTA Format



Abstract Syntax Notation: ASN.1



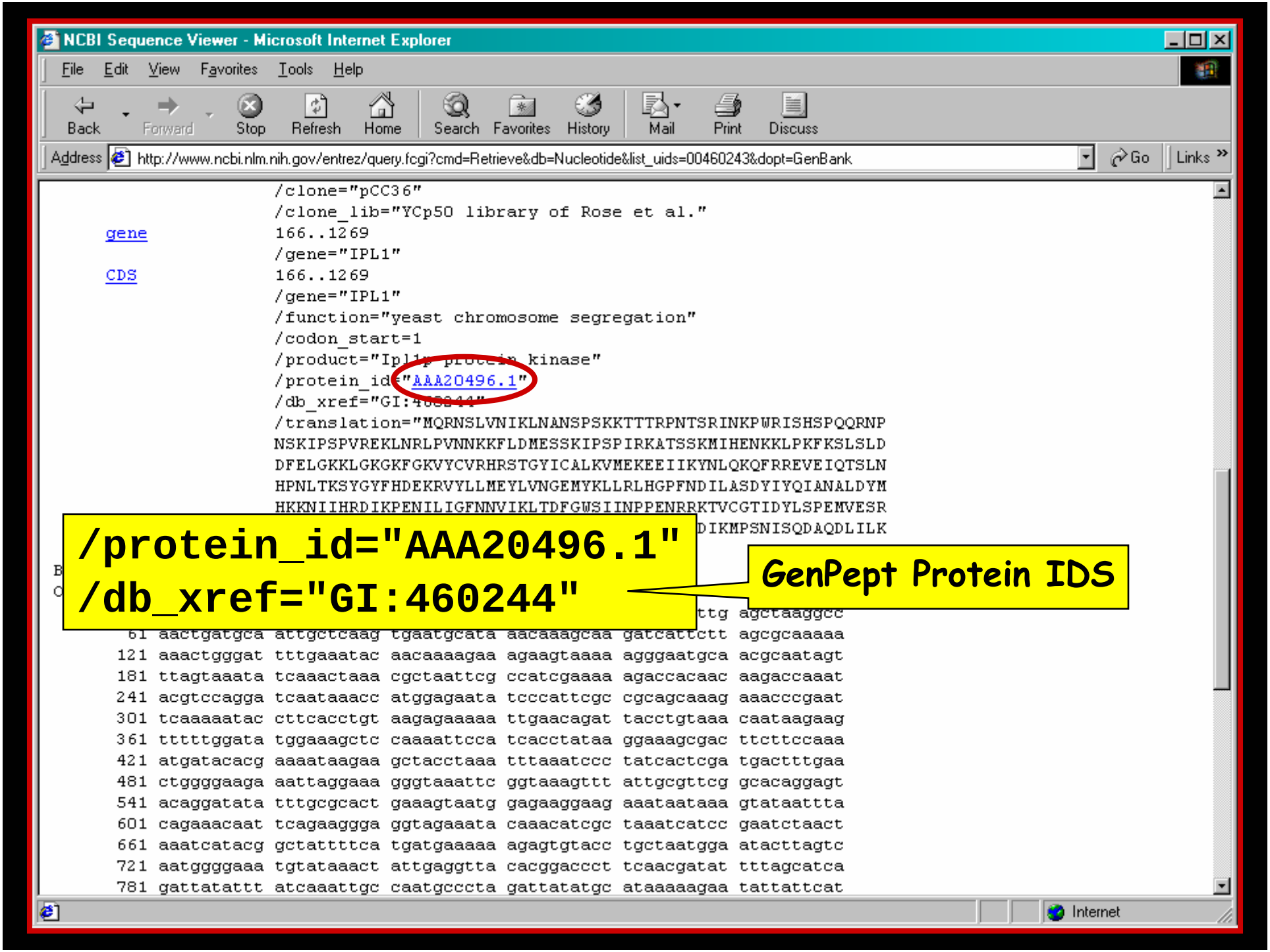
NCBI Toolbox

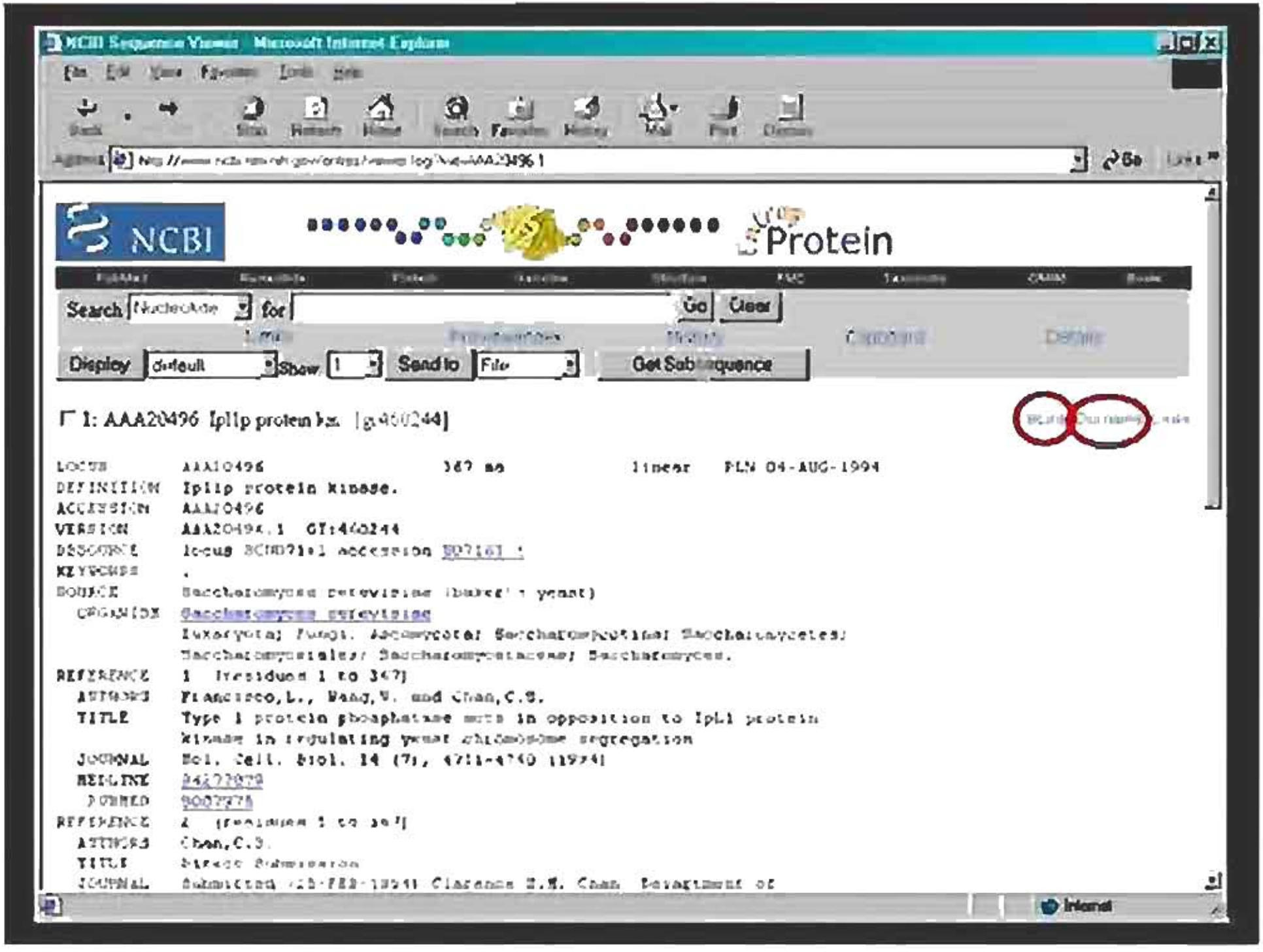
```
 /**************************************************************************
*
*   asn2ff.c
*         convert an ASN.1 entry to flat file format, using the FFPrintArray.
*
***************
#include <accent
#include "asn2ff
#include "asn2ff
#include "ffprin
#include <subuti
#include <objall
#include <objcod
#include <lsqfet
#include <explor

#ifdef ENABLE_ID1
#include <accid1
#endif

FILE *fpl;

Args myargs[] = {
          {"Input asnfile in binary mode","F",NULL,NULL,TRUE,'b',ARG_BOOLEAN,0.0,0,NULL},
          {"Output Filename","stdout", NULL,NULL,TRUE,'o',ARG_FILE_OUT,0.0,0,NULL},
          {"Show Sequence?","T", NULL ,NULL ,TRUE,'h',ARG_BOOLEAN,0.0,0,NULL},
```

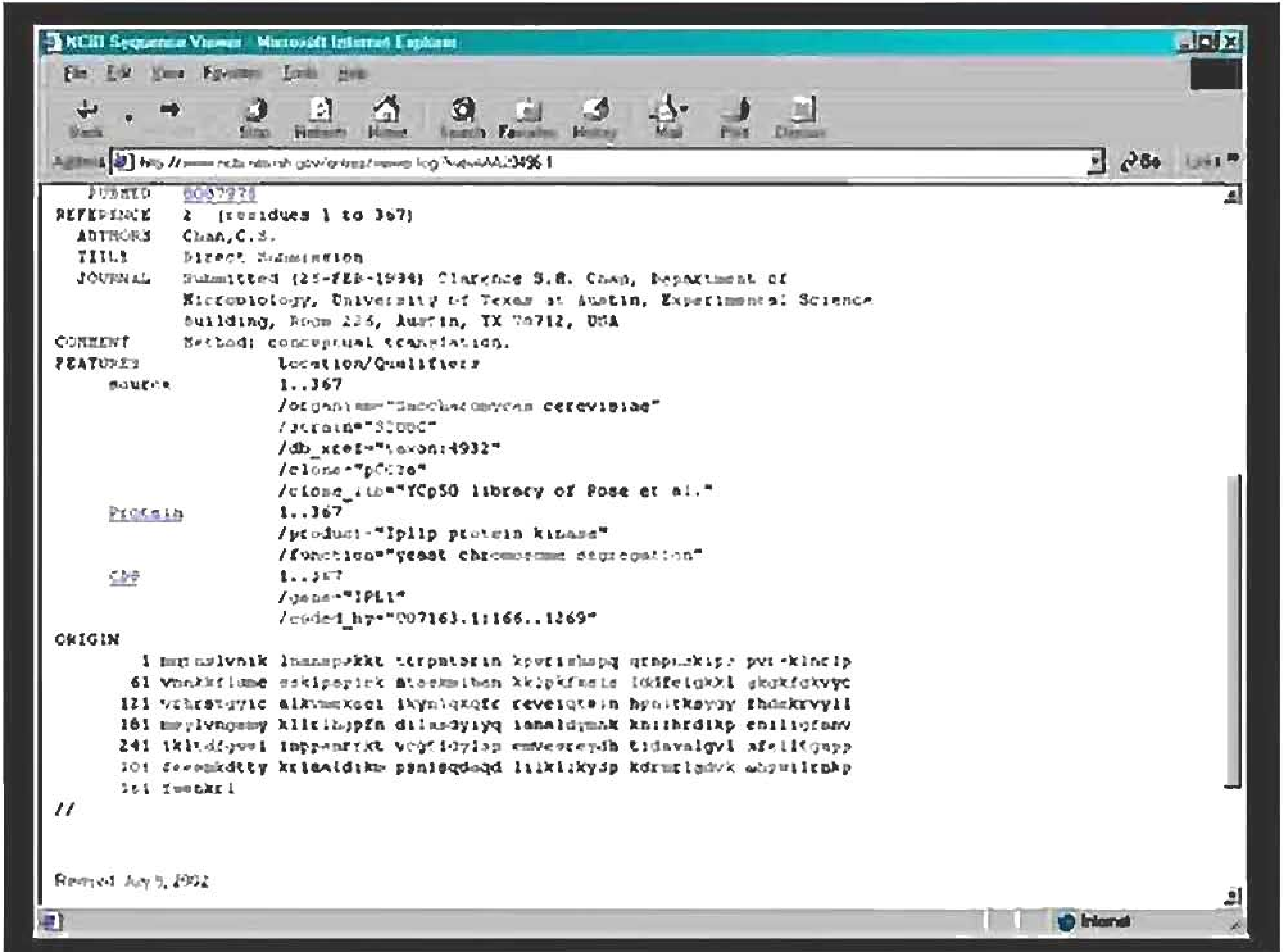
Toolbox Sources

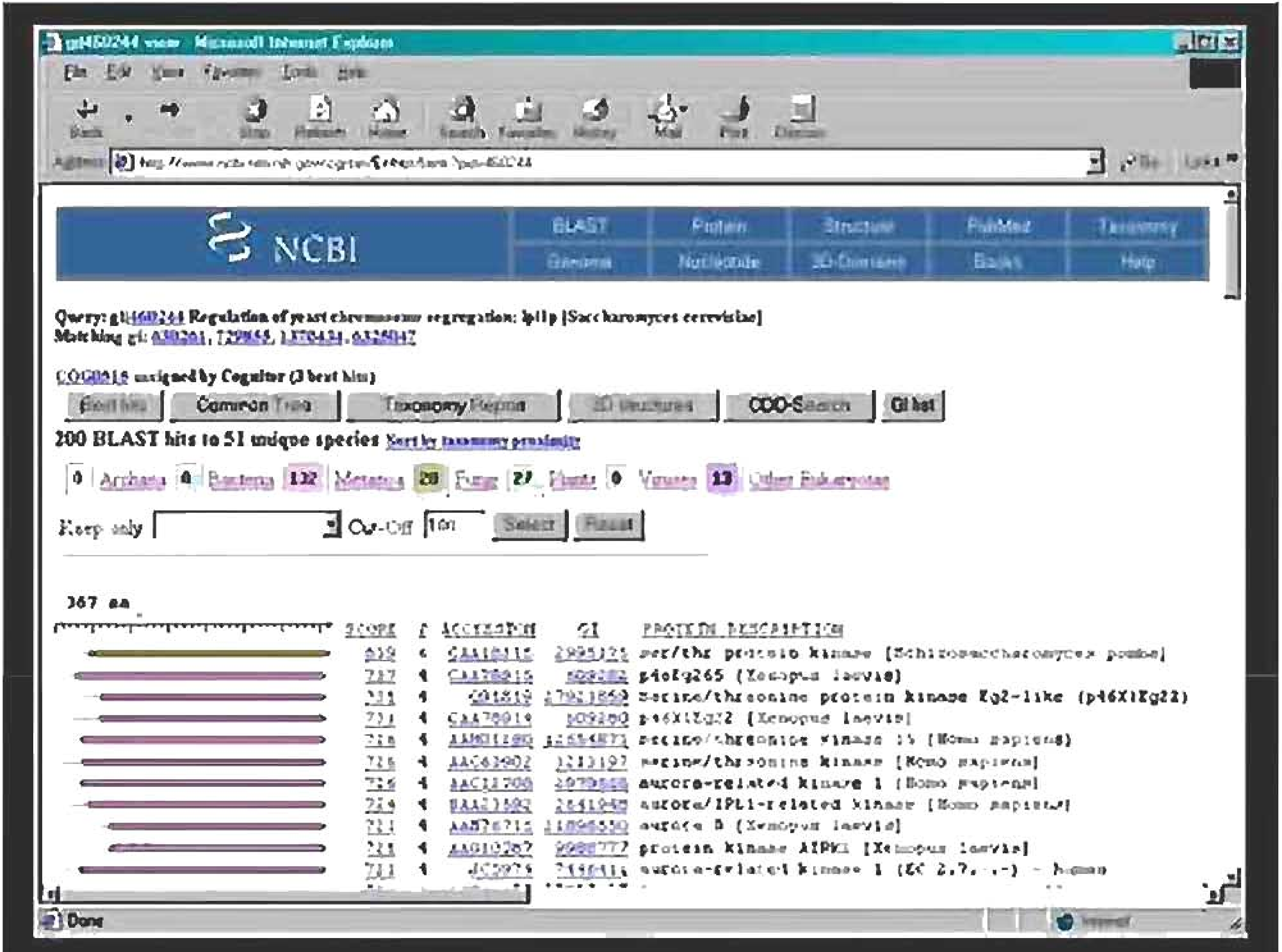
```
ftp> open ftp.ncbi.nih.gov
.
.
ftp> cd toolbox
ftp> cd ncbi_tools
```

ftp://ftp.ncbi.nlm.gov/toolbox/ncbi_tools

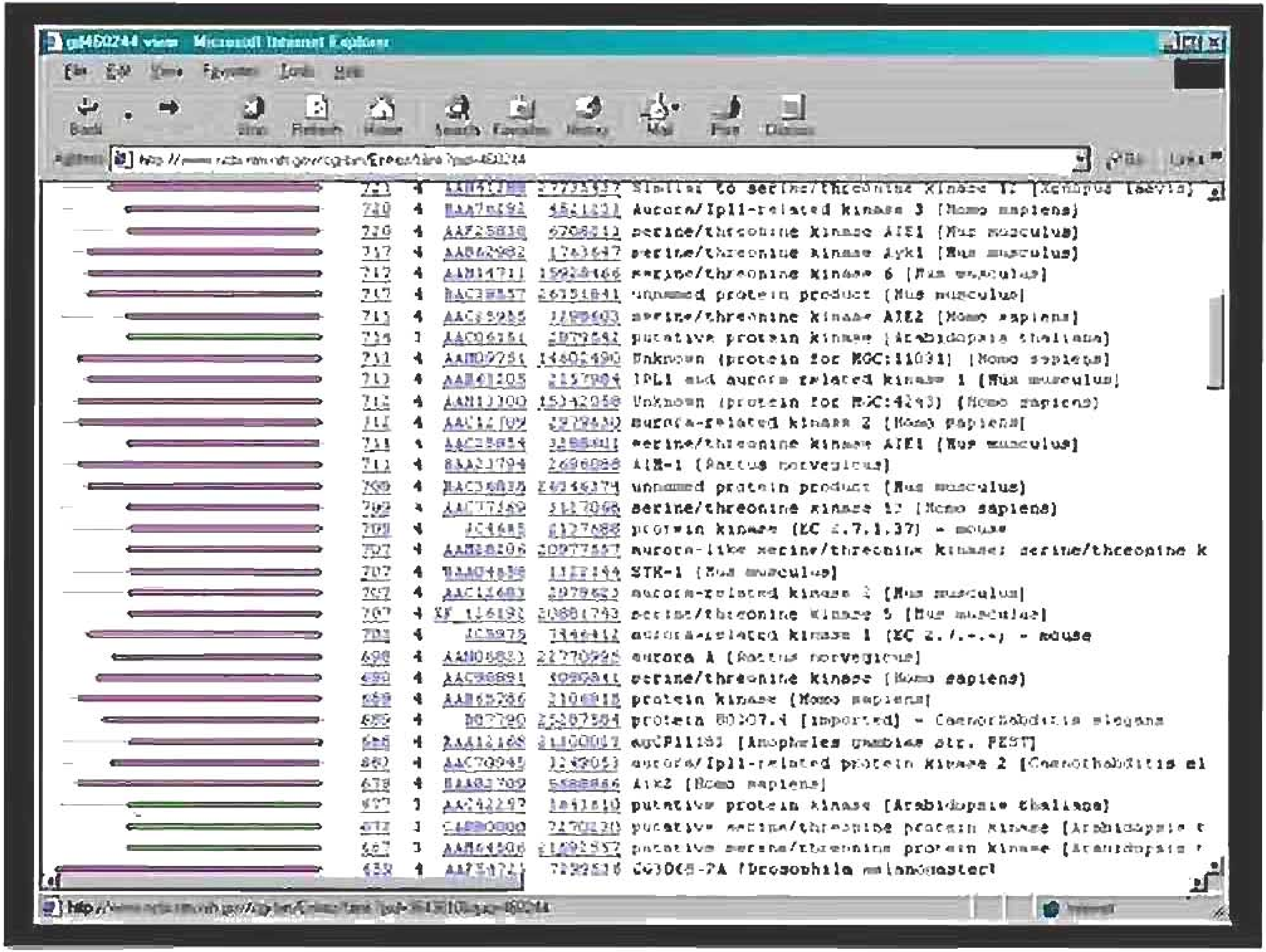

NCBI Sequence Viewer - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites History Mail Print Discuss
Address http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=Nucleotide&list_uids=00460243&dopt=GenBank
Go Links
/clone="pCC36"
/clone_lib="YCp50 library of Rose et al."
gene 166..1269
/gene="IPL1"
CDS 166..1269
/gene="IPL1"
/function="yeast chromosome segregation"
/codon_start=1
/product="Ipl1p protein kinase"
/protein_id="AAA20496.1"
/db_xref="GI:460244"
/translation="MQRNSLVNIKLNANSPSKKTTTRPNTSRINKPWRISHSPQQRNP
NSKIPSPVREKLNRLPVNNKKFLDMESSKIPSPIRKATSSKMIHENKKLPKFKSLSLD
DFELGKKLGKGKFGKVYCVRHRSTGYICALKVMEKEEIIKYNLQKQFRREVEIQTSLN
HPNLTKSYGYFHDEKRVYLLMEYLVNGEMYKLLRLHGPFNDILASDYIYQIANALDYM
HKKNIIHRDIKPENILIGFNNVIKLTDFGWSIINPPENRRKTVCGTIDYLSPEMVESR
/protein_id="AAA20496.1"
/db_xref="GI:460244"
GenPept Protein IDS
61 aactgatgca attgctcaag tgaatgcata aacaaagcaa gatcattctt agcgcaaaaa
121 aaactgggat tttgaaatac aacaaaagaa agaagtaaaa agggaatgca acgcaatagt
181 ttagtaaata tcaaactaaa cgctaattcg ccatcgaaaa agaccacaac aagaccaaat
241 acgtccagga tcaataaacc atggagaata tcccattcgc cgcagcaaag aaacccgaat
301 tcaaaaatac cttcacctgt aagagaaaaa ttgaacagat tacctgtaaa caataagaag
361 tttttggata tggaaagctc caaaattcca tcacctataa ggaaagcgac ttcttccaaa
421 atgatacacg aaaataagaa gctacctaaa tttaaatccc tatcactcga tgactttgaa
481 ctggggaaga aattaggaaa gggtaaattc ggtaaagttt attgcgttcg gcacaggagt
541 acaggatata tttgcgcact gaaagtaatg gagaaggaag aaataataaa gtataattta
601 cagaaacaat tcagaaggga ggtagaaata caaacatcgc taaatcatcc gaatctaact
661 aaatcatacg gctattttca tgatgaaaaa agagtgtacc tgctaatgga atacttagtc
721 aatggggaaa tgtataaact attgaggtta cacggaccct tcaacgatat tttagcatca
781 gattatattt atcaaattgc caatgcccta gattatatgc ataaaaagaa tattattcat
Internet


NCBI
Protein
Search
for
Go
Clear
Display
default
Show
1
Send to
File
Get Subsequence
1: AAA20496 Ipl1p protein ki... [g460244]
LOCUS AAA20496 367 aa linear PLN 04-AUG-1994
DEFINITION Ipl1p protein kinase.
ACCESSION AAA20496
VERSION AAA20496.1 GI:460244
KEYWORDS .
SOURCE Saccharomyces cerevisiae (baker's yeast)
ORGANISM Saccharomyces cerevisiae
Eukaryota; Fungi; Ascomycota; Saccharomycotina; Saccharomycetes;
Saccharomycetales; Saccharomycetaceae; Saccharomyces.
REFERENCE 1 (residues 1 to 367)
AUTHORS Francisco,L., Wang,W. and Chan,C.S.
TITLE Type 1 protein phosphatase acts in opposition to IpL1 protein
kinase in regulating yeast chromosome segregation
JOURNAL Mol. Cell. Biol. 14 (7), 4731-4740 (1994)
MEDLINE 94327879
PUBMED 8007978
REFERENCE 2 (residues 1 to 367)
AUTHORS Chan,C.S.
TITLE Direct Submission
JOURNAL Submitted (28-FEB-1994) Clarence S.M. Chan, Department of
Internet

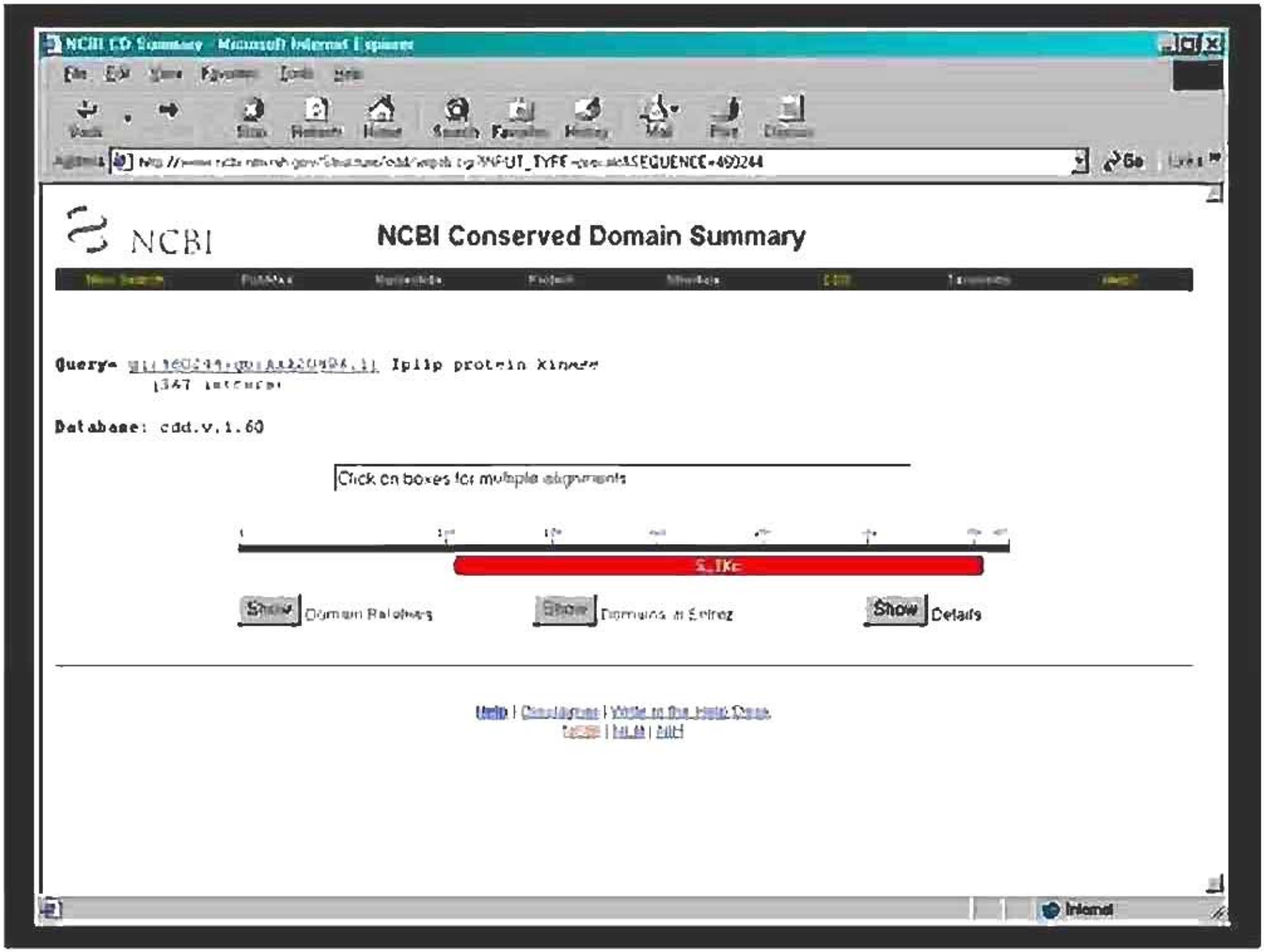

REFERENCE 2 (residues 1 to 367)
AUTHORS Chan,C.S.
TITLE Direct Submission
JOURNAL Submitted (25-FEB-1994) Clarence S.M. Chan, Department of Microbiology, University of Texas at Austin, Experimental Science Building, Austin, TX 78712, USA
COMMENT Method: conceptual translation.
FEATURES Location/Qualifiers
source 1..367
/organism="Saccharomyces cerevisiae"
/db_xref="taxon:4932"
/clone_lib="YCp50 library of Rose et al."
Protein 1..367
/product="Ipl1p protein kinase"
/function="yeast chromosome segregation"
CDS
/gene="IPL1"
ORIGIN
//

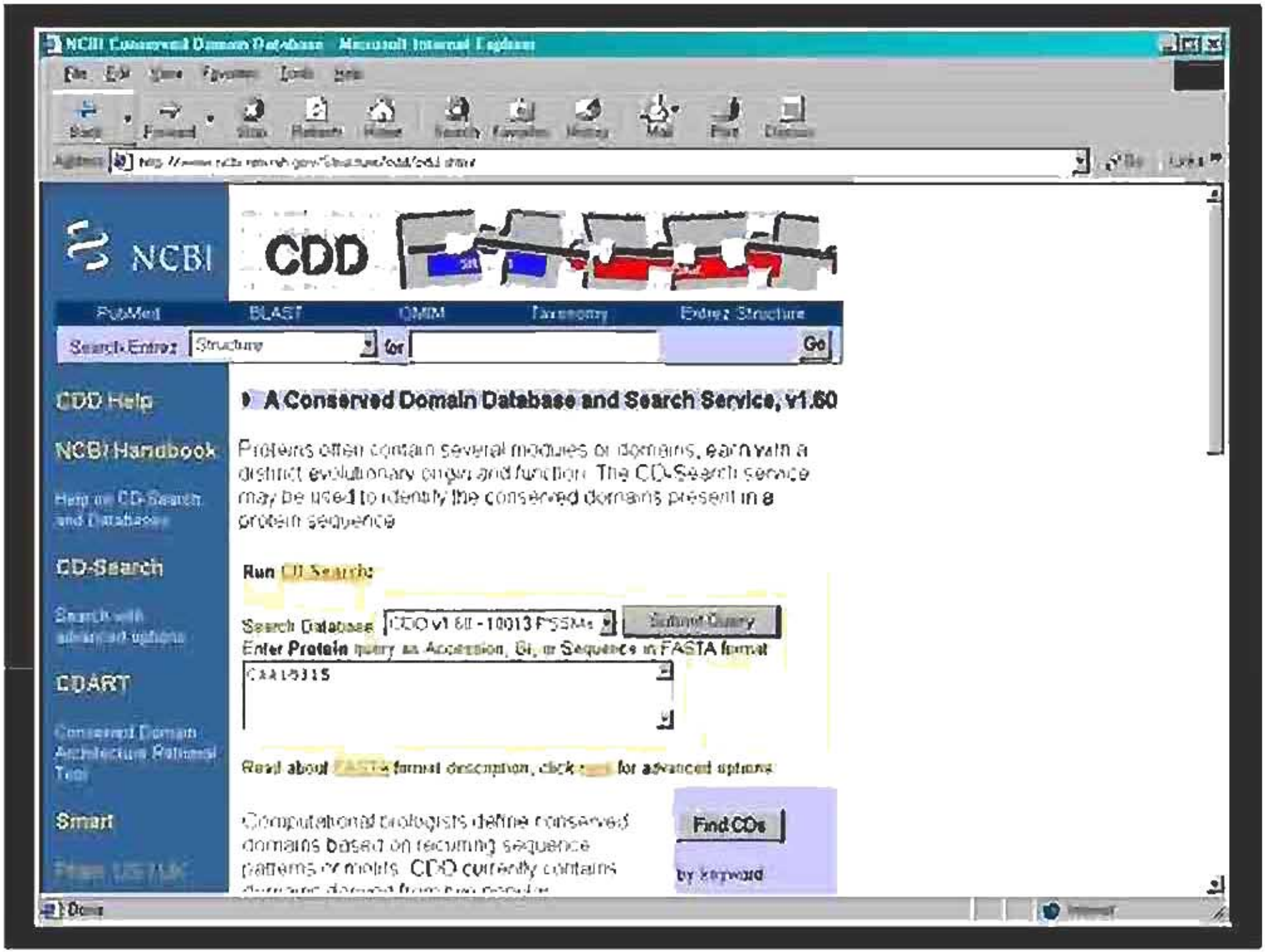


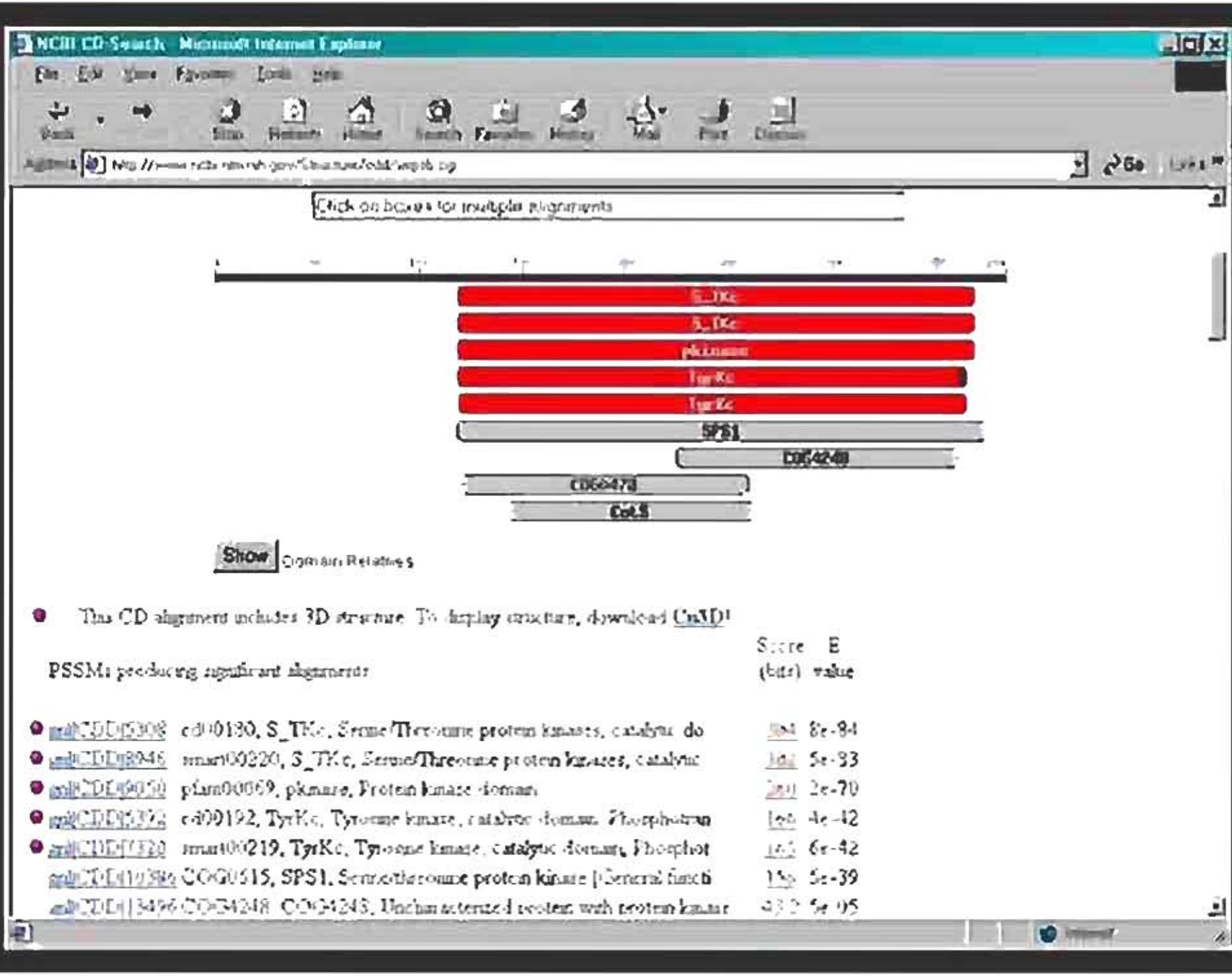
NCBI
BLAST
Protein
Structure
PubMed
Taxonomy
Nucleotide
Books
Help
Query: gi|460244 Regulation of yeast chromosome segregation; Ipl1p [Saccharomyces cerevisiae]
Common Tree
Taxonomy Reports
COG-Search
GI list
200 BLAST hits to 51 unique species
Archaea
Bacteria
Metazoa
Fungi
Plants
Viruses
Other Eukaryotes
Keep only
Cut-Off
Select
Reset
SCORE
ACCESSION
GI
PROTEIN DESCRIPTION
ser/thr protein kinase [Schizosaccharomyces pombe]
serine/threonine protein kinase Eg2-like (p46XlEg22)
serine/threonine kinase [Homo sapiens]
aurora-related kinase 1 [Homo sapiens]
aurora/IPL1-related kinase [Homo sapiens]
aurora B [Xenopus laevis]
Done

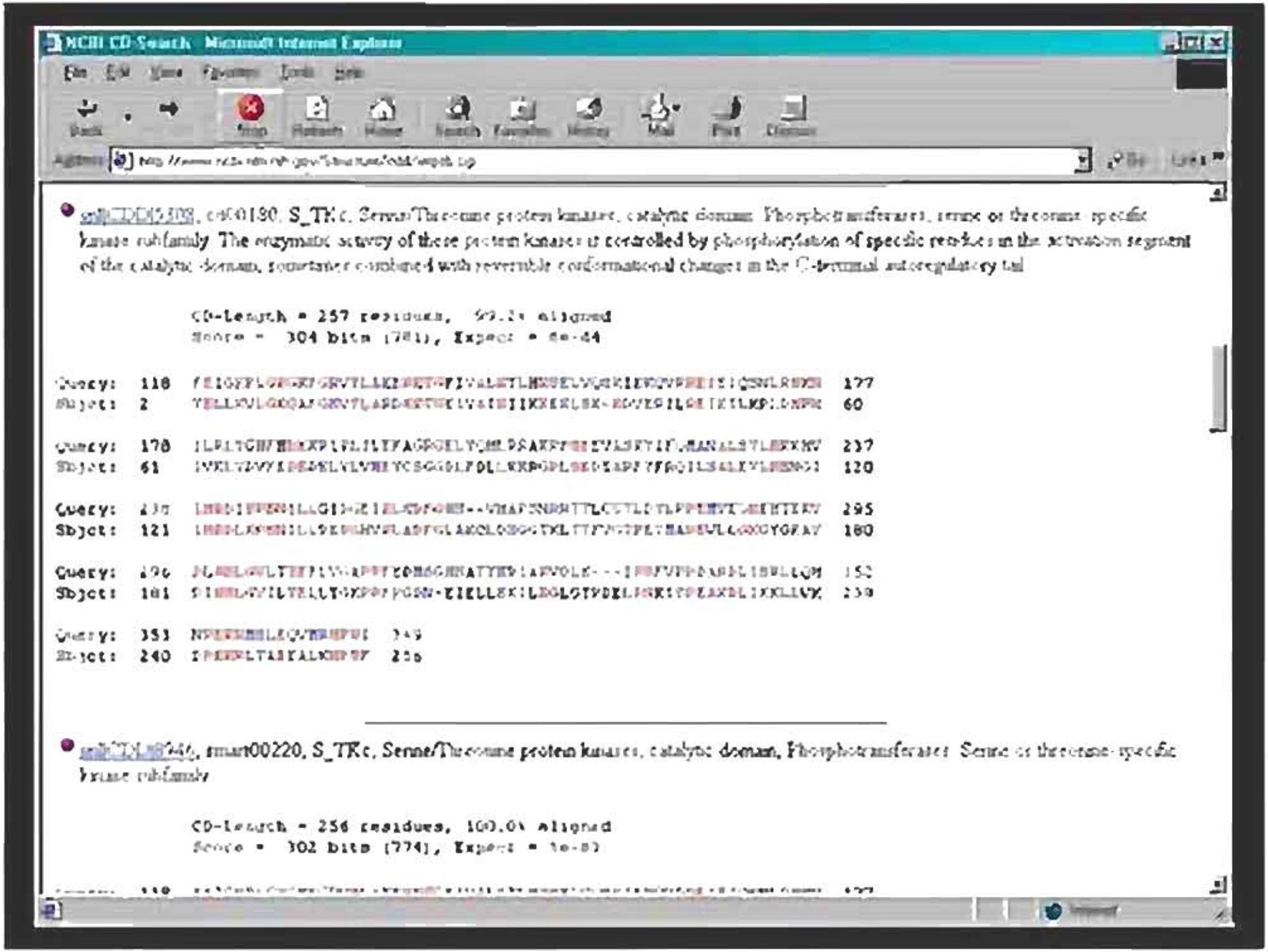


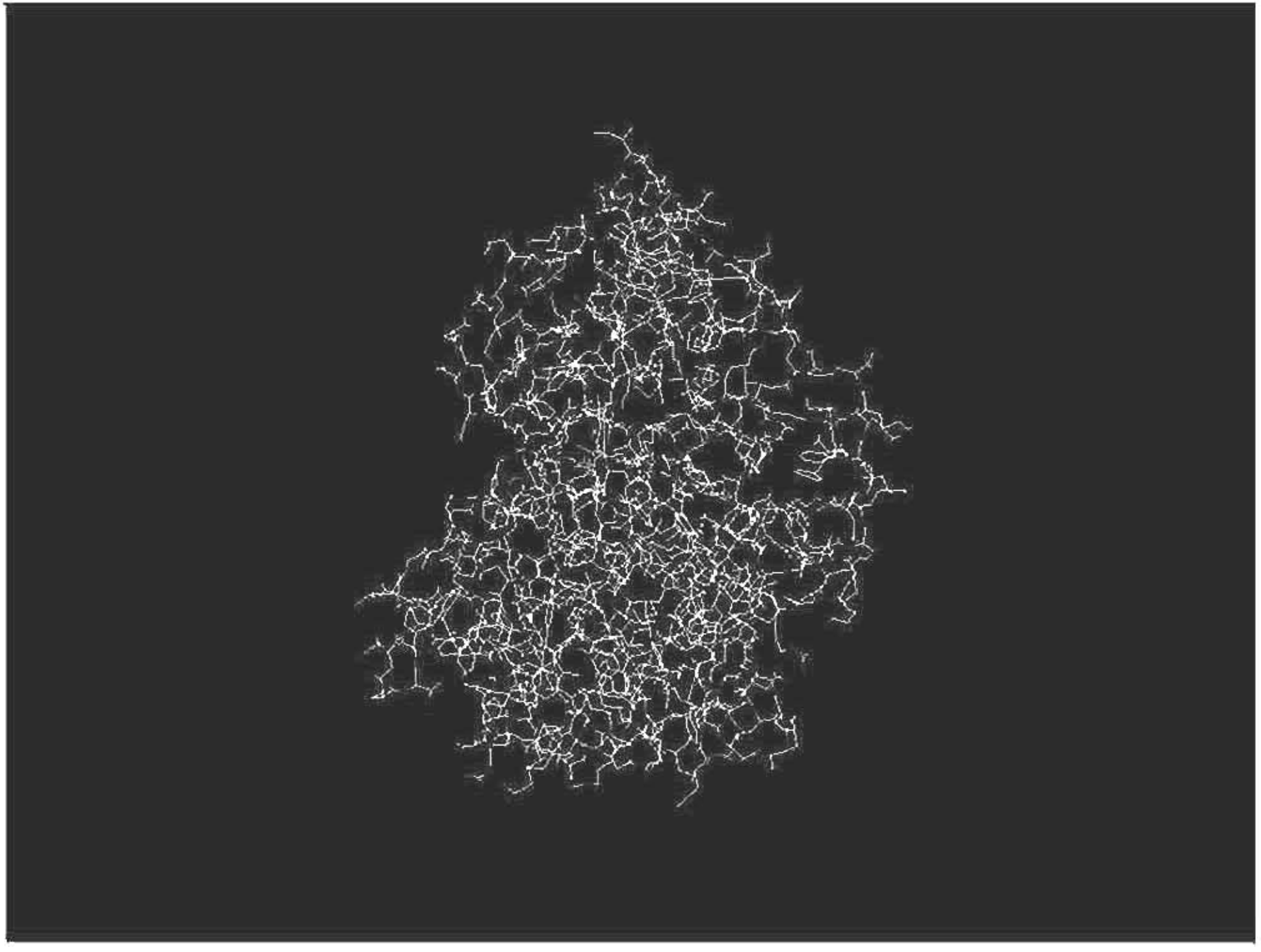
Microsoft Internet Explorer
Back
Stop
Refresh
Home
Search
Favorites
History
Mail
Print
Aurora/Ipl1-related kinase 3 (Homo sapiens)
serine/threonine kinase AIE1 (Mus musculus)
serine/threonine kinase Ayk1 (Mus musculus)
serine/threonine kinase 6 (Mus musculus)
unnamed protein product (Mus musculus)
serine/threonine kinase AIE2 (Homo sapiens)
putative protein kinase (Arabidopsis thaliana)
Unknown (protein for MGC:11031) (Homo sapiens)
IPL1 and aurora related kinase 1 (Mus musculus)
aurora-related kinase 2 (Homo sapiens)
serine/threonine kinase AIE1 (Mus musculus)
unnamed protein product (Mus musculus)
serine/threonine kinase 12 (Homo sapiens)
protein kinase (EC 2.7.1.37) - mouse
STK-1 (Mus musculus)
serine/threonine kinase 5 (Mus musculus)
aurora-related kinase 1 (EC 2.7.-.-) - mouse
aurora A (Rattus norvegicus)
serine/threonine kinase (Homo sapiens)
protein kinase (Homo sapiens)
agCP11193 [Anopheles gambiae str. PEST]
aurora/Ipl1-related protein kinase 2 [Caenorhabditis el
AIK2 (Homo sapiens)
putative protein kinase [Arabidopsis thaliana]
CG3068-PA [Drosophila melanogaster]

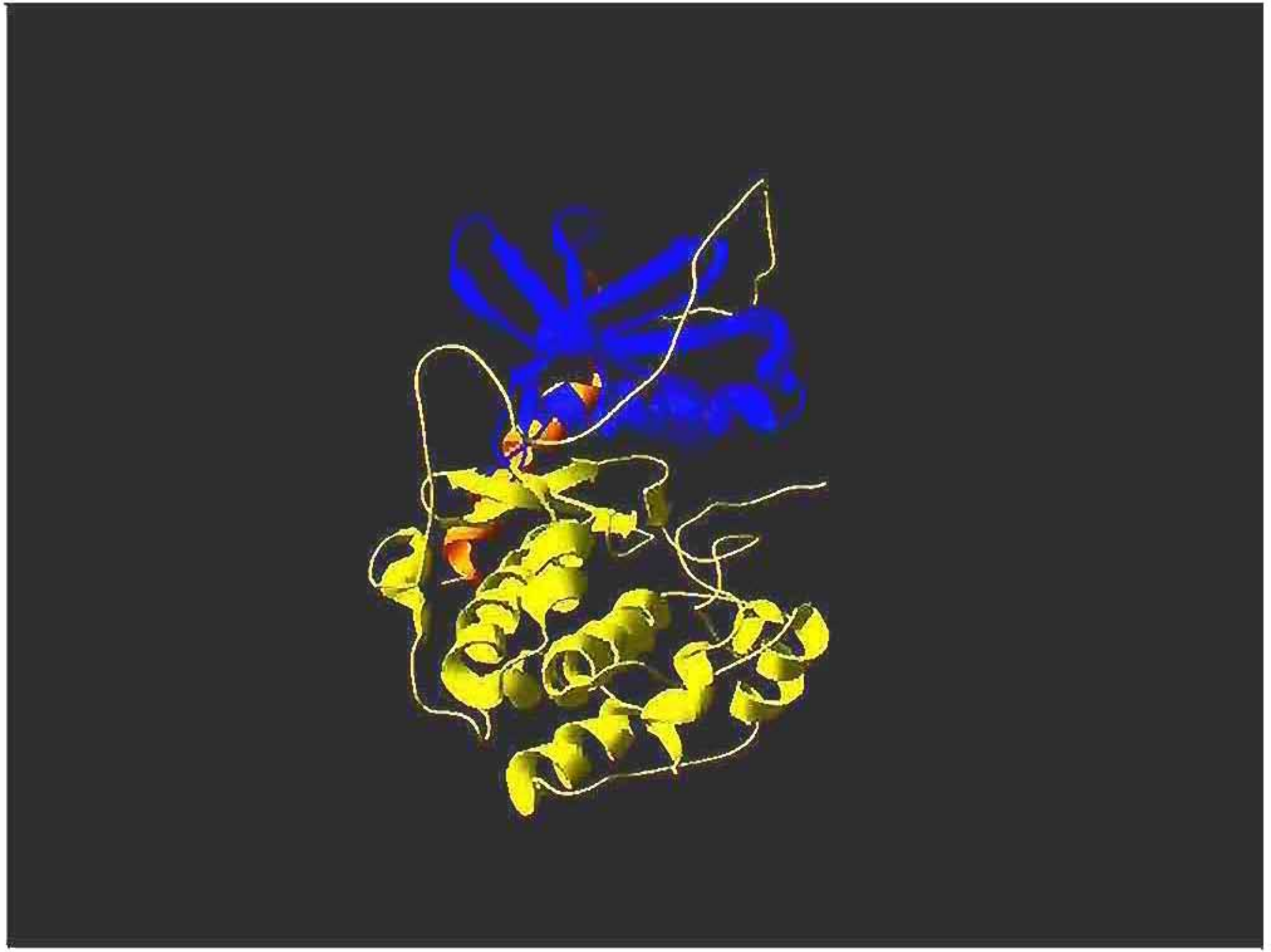

NCBI
NCBI Conserved Domain Summary
Query=
1347 letters
Database: cdd.v.1.60
Click on boxes for multiple alignments
S_TKc
Show
Domain Relatives
Show
Domains in Entrez
Show
Details
Help
Internet


NCBI
CDD
PubMed
BLAST
OMIM
Taxonomy
Entrez Structure
Search Entrez
Structure
for
Go
CDD Help
NCBI Handbook
Help on CD-Search and Databases
CD-Search
Search with advanced options
CDART
Conserved Domain Architecture Retrieval Tool
Smart
A Conserved Domain Database and Search Service, v1.60
Proteins often contain several modules or domains, each with a distinct evolutionary origin and function. The CD-Search service may be used to identify the conserved domains present in a protein sequence
Run CD-Search
Search Database
Enter Protein query as Accession, GI, or Sequence in FASTA format
Read about FASTA format description, click here for advanced options
Computational biologists define conserved domains based on recurring sequence patterns or motifs. CDD currently contains
Find CDs
by keyword


NCBI CD-Search Microsoft Internet Explorer
Click on boxes for multiple alignments
S_TKc
S_TKc
pkinase
TyrKc
TyrKc
SPS1
COG4248
COG5478
COG5
Show
Domain Relatives
This CD alignment includes 3D structure. To display structure, download Cn3D!
PSSMs producing significant alignments
Score (bits)
E value
cd00180, S_TKc, Serine/Threonine protein kinases, catalytic do
8e-84
smart00220, S_TKc, Serine/Threonine protein kinases, catalytic
5e-83
pfam00069, pkinase, Protein kinase domain
2e-70
cd00192, TyrKc, Tyrosine kinase, catalytic domain. Phosphotran
4e-42
smart00219, TyrKc, Tyrosine kinase, catalytic domain, Phosphot
6e-42
COG0515, SPS1, Serine/threonine protein kinase [General functi
5e-39
COG4248, Uncharacterized protein with protein kinase


NCBI CD-Search - Microsoft Internet Explorer
cd00180, S_TKc, Serine/Threonine protein kinases, catalytic domain; Phosphotransferases; serine or threonine-specific kinase subfamily. The enzymatic activity of these protein kinases is controlled by phosphorylation of specific residues in the activation segment of the catalytic domain, sometimes combined with reversible conformational changes in the C-terminal autoregulatory tail
CD-Length = 257 residues, 99.2% aligned
Score = 304 bits (781), Expect = 6e-84
Query: 118 177
Sbjct: 2 60
Query: 178 237
Sbjct: 61 120
Query: 238 295
Sbjct: 121 180
Query: 296 352
Sbjct: 181 238
Query: 353 349
Sbjct: 240 286
smart00220, S_TKc, Serine/Threonine protein kinases, catalytic domain; Phosphotransferases; Serine or threonine-specific kinase subfamily
CD-Length = 256 residues, 100.0% aligned
Score = 302 bits (774), Expect = 1e-83






BREAK

Why align sequences?

- Functional predictions based on identifying homologues.

Assumes:

conservation of sequence  conservation of function

BUT: Function carried out at level of proteins, i.e. 3-D structure

Sequence conservation carried out at level of DNA 1-D sequence

Implicit Assumption of Evolution in Models of Sequence Homology

Assume:

Sequence conservation  Structure conservation

---

Note that the converse is NOT necessarily true!

Structure conservation  Sequence conservation

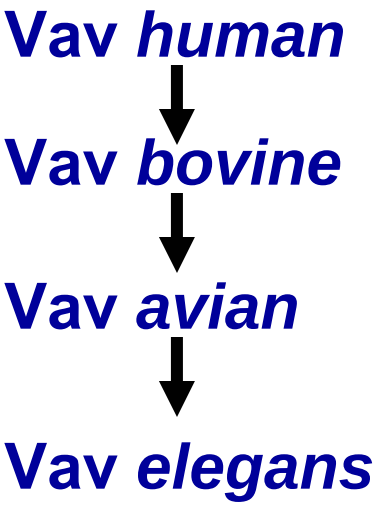
Definitions

- **Homologue:**

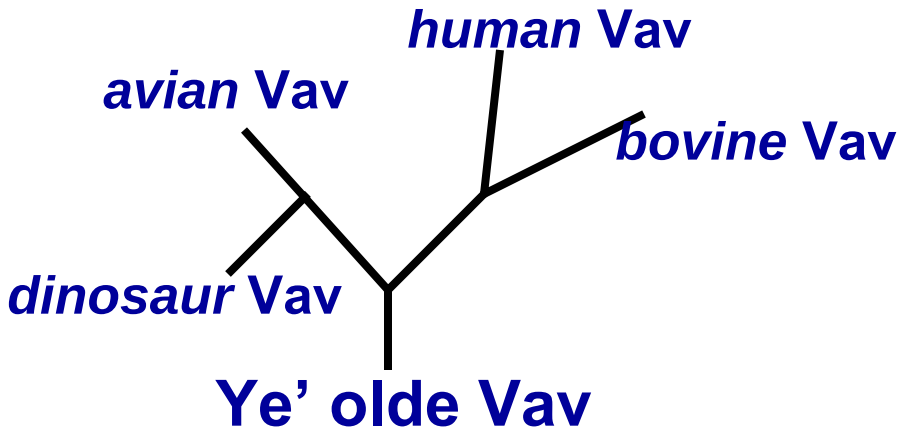
a relatively non-specific term (meaningless?) that conveys the idea that two sequences are somehow related

- **Orthologue:**

Ortho = (*greek*) straight….implies direct descent, 1 ancestor



-or-

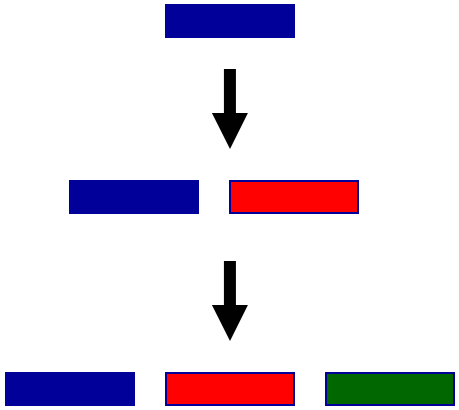


Definitions

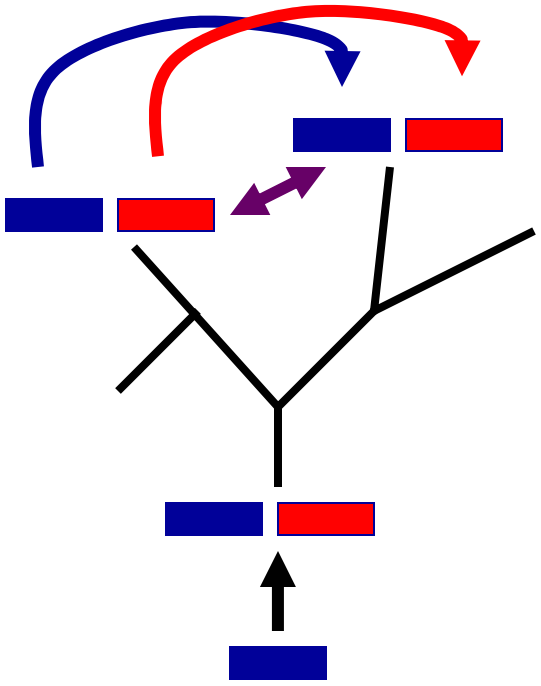
- **Paralogue:**

Para = (*greek*) along side of….implies some type of gene duplication event

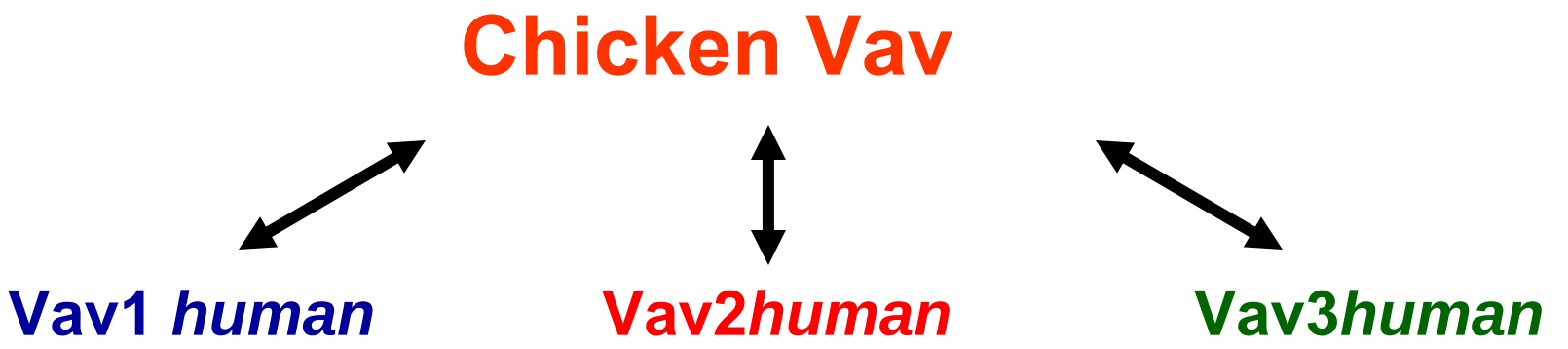
Vav1 *human* ⟷ **Vav2*human*** ⟷ **Vav3*human***



-or-



Complex problem



Orthologue or Paralogue?

Need to know evolutionary relationships

Can try and get these from alignments as well…

Alignments- Good Bad and Ugly

HgbA-human

```
GSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKL
:+ +::+::::::+ :++++++::+:++ ++++++::+::+ ::
GNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKL
```

HgbB-human

Alignments- Good Bad and Ugly

SPURIOUS ALIGNMENT

HgbA-human

```
GSAQVKGHGKKVADALTNAVAHVDDMPNALSALSD----LHAHKL
:::+ ++: +  ++:+:  ++ :+ :+  : +:+ :    ++::+
GSGYLVGDSLTFVDLL--VAQHTADLLAANAALLDEFPQFKAHQE
```

Nematode glutathione S-transferase

HgbA-human

```
GSAQVKGHGKKVADALTNAVAHV---D--DMPNALSALSDLHAHKL
++ ++++:+ ::+ ++ +:++ +            +:  :+ +:+ :
NNPELQAHAGKVFKLVYEAAIQLQVTGVVVTDATLKNLGSVHVSKG
```

Leghemoglobin, yellow lupin

Alignments

- **Types:**
  - **Local**
  - **Global**
  - **Ungapped**
  - **Gapped (2 types- linear, affine)**
- **Methods:**
  - **Dot matrix**
  - **Dynamic Programming**
  - **Word, k-tup**

Simplest alignment representation – Dot plot

Model: Need a metric of similarity between amino acid pairs

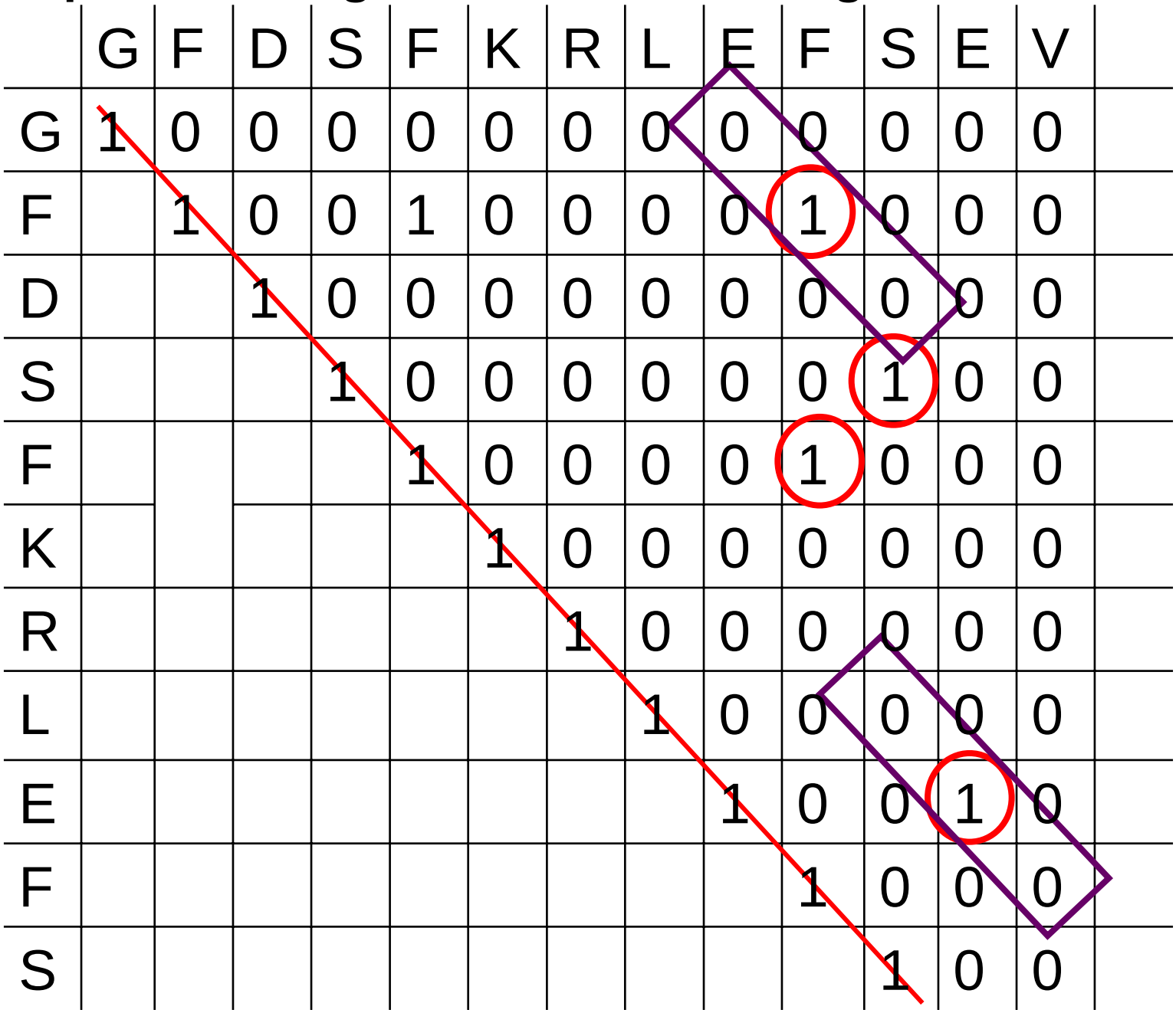
Simplest metric – unitary matrix, identity matrix

	A	C	D	E	F	G	H	I	K
A	1	0	0	0	0	0	0	0	0
C		1	0	0	0	0	0	0	0
D			1	0	0	0	0	0	0
E				1	0	0	0	0	0
F					1	0	0	0	0
G						1	0	0	0
H							1	0	0
I								1	0
K									1

Example – self alignment

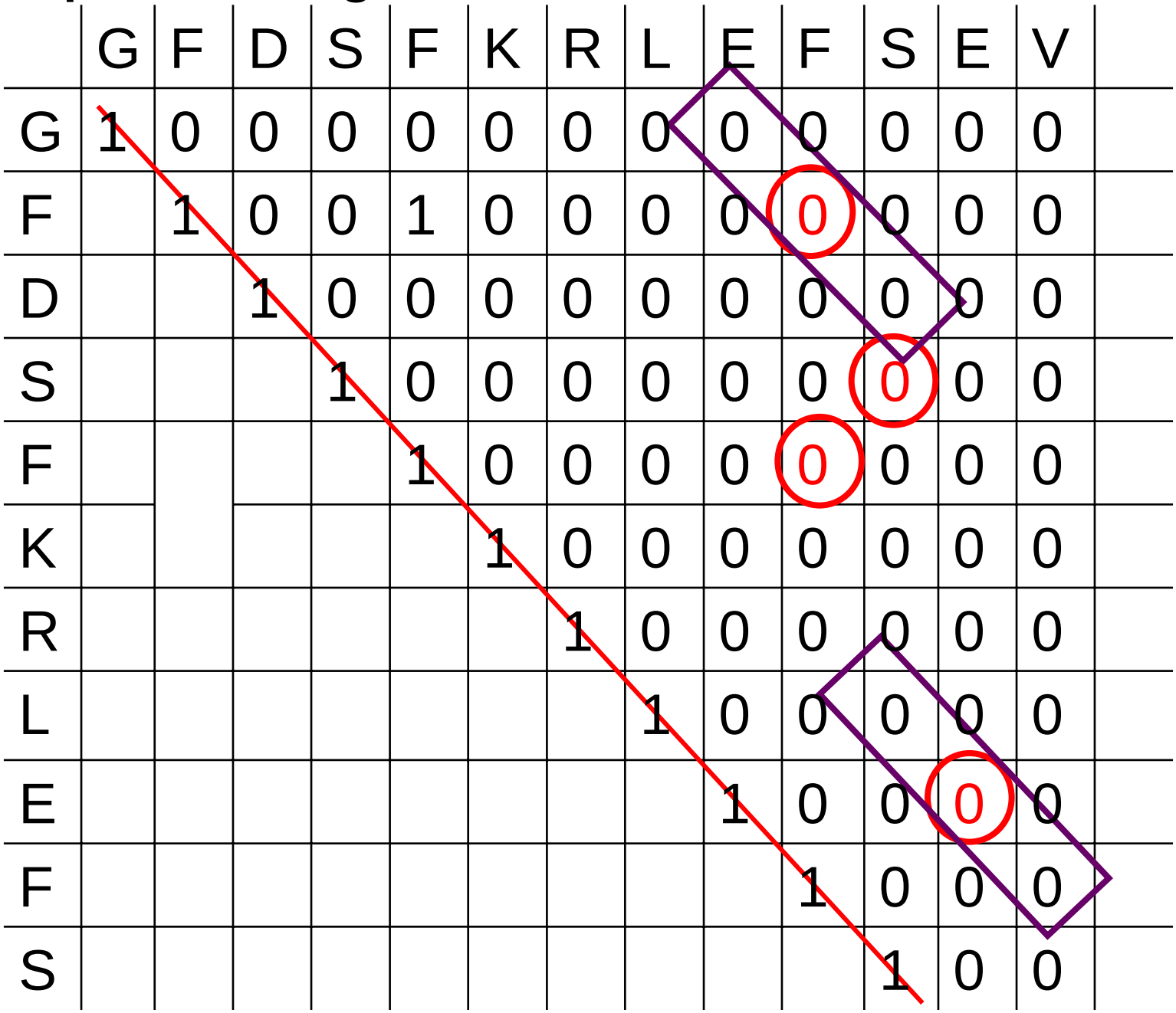
	G	F	D	S	F	K	R	L	E	F	S	E	V	
G	1	0	0	0	0	0	0	0	0	0	0	0	0	
F		1	0	0	1	0	0	0	0	1	0	0	0	
D			1	0	0	0	0	0	0	0	0	0	0	
S				1	0	0	0	0	0	0	1	0	0	
F					1	0	0	0	0	1	0	0	0	
K						1	0	0	0	0	0	0	0	
R							1	0	0	0	0	0	0	
L								1	0	0	0	0	0	
E									1	0	0	1	0	
F										1	0	0	0	
S											1	0	0	

Example – self alignment…..with sliding window



	G	F	D	S	F	K	R	L	E	F	S	E	V	
G	1	0	0	0	0	0	0	0	0	0	0	0	0	
F		1	0	0	1	0	0	0	0	1	0	0	0	
D			1	0	0	0	0	0	0	0	0	0	0	
S				1	0	0	0	0	0	0	1	0	0	
F					1	0	0	0	0	1	0	0	0	
K						1	0	0	0	0	0	0	0	
R							1	0	0	0	0	0	0	
L								1	0	0	0	0	0	
E									1	0	0	1	0	
F										1	0	0	0	
S											1	0	0	

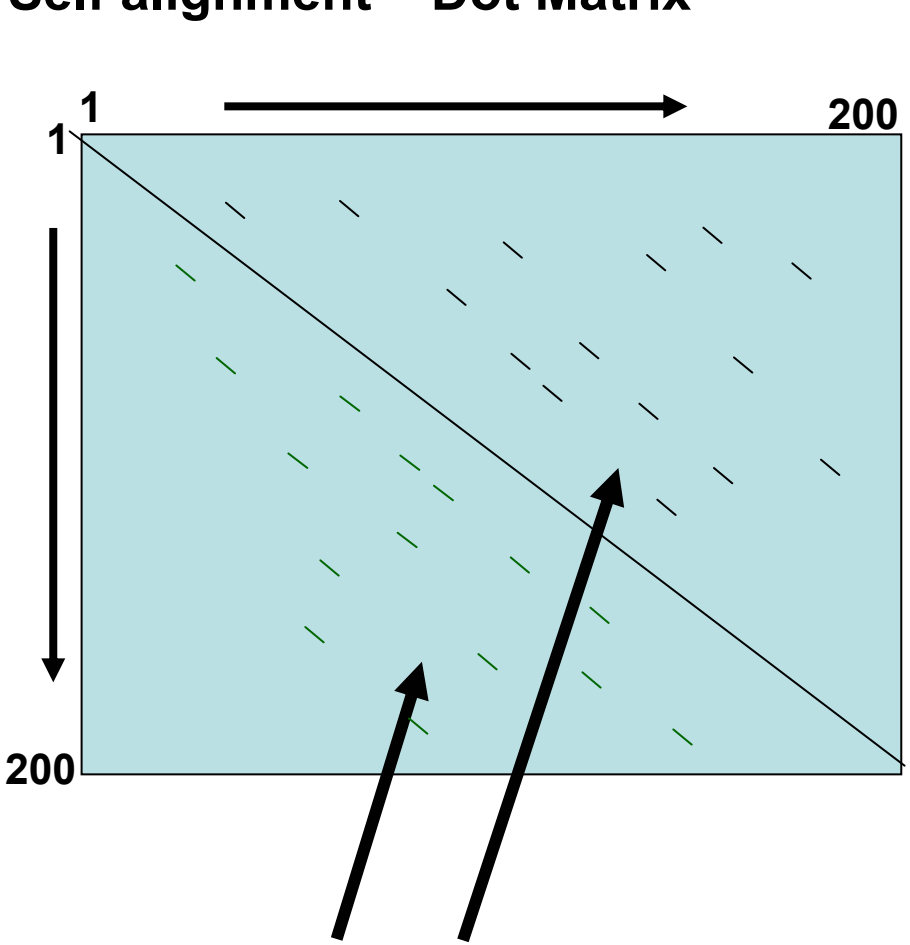
Example – self alignment



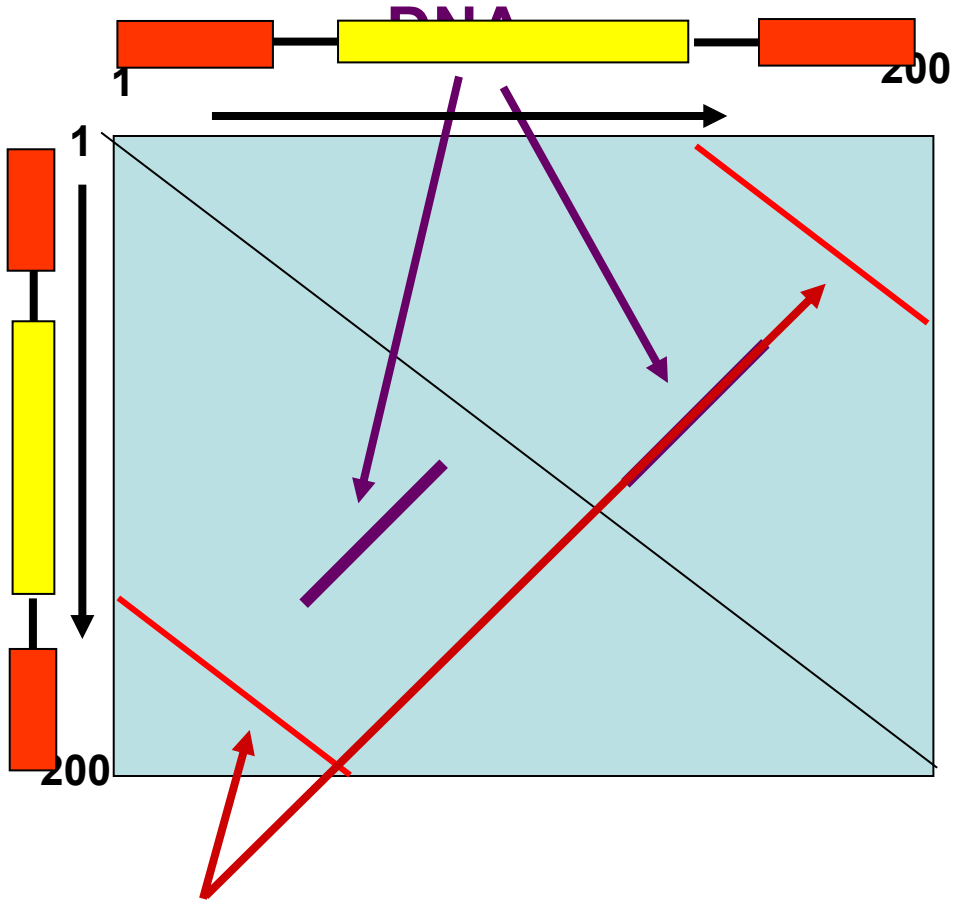
	G	F	D	S	F	K	R	L	E	F	S	E	V	
G	1	0	0	0	0	0	0	0	0	0	0	0	0	
F		1	0	0	1	0	0	0	0	0	0	0	0	
D			1	0	0	0	0	0	0	0	0	0	0	
S				1	0	0	0	0	0	0	0	0	0	
F					1	0	0	0	0	0	0	0	0	
K						1	0	0	0	0	0	0	0	
R							1	0	0	0	0	0	0	
L								1	0	0	0	0	0	
E									1	0	0	0	0	
F										1	0	0	0	
S											1	0	0	

Simple alignments

Self alignment – Dot Matrix



symmetric

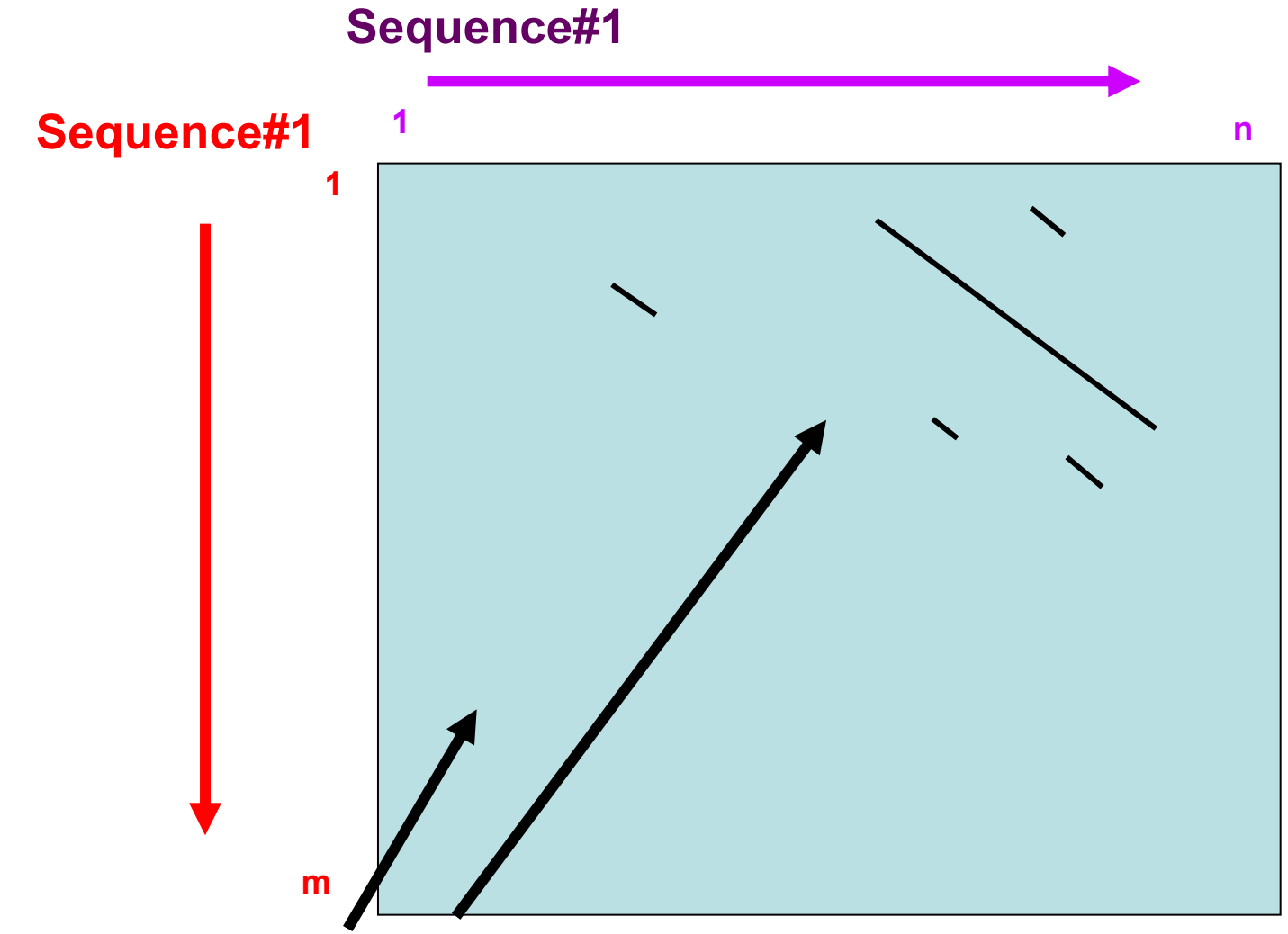


Now align two different sequences

* **Consider other similarity matrices besides identity**….

- ***Chemical similarity*** – binary decision
- ***Amino acid conservation*** in aligned protein families – min. similarity score (+/- window)
- ***Average*** of multiple scoring systems

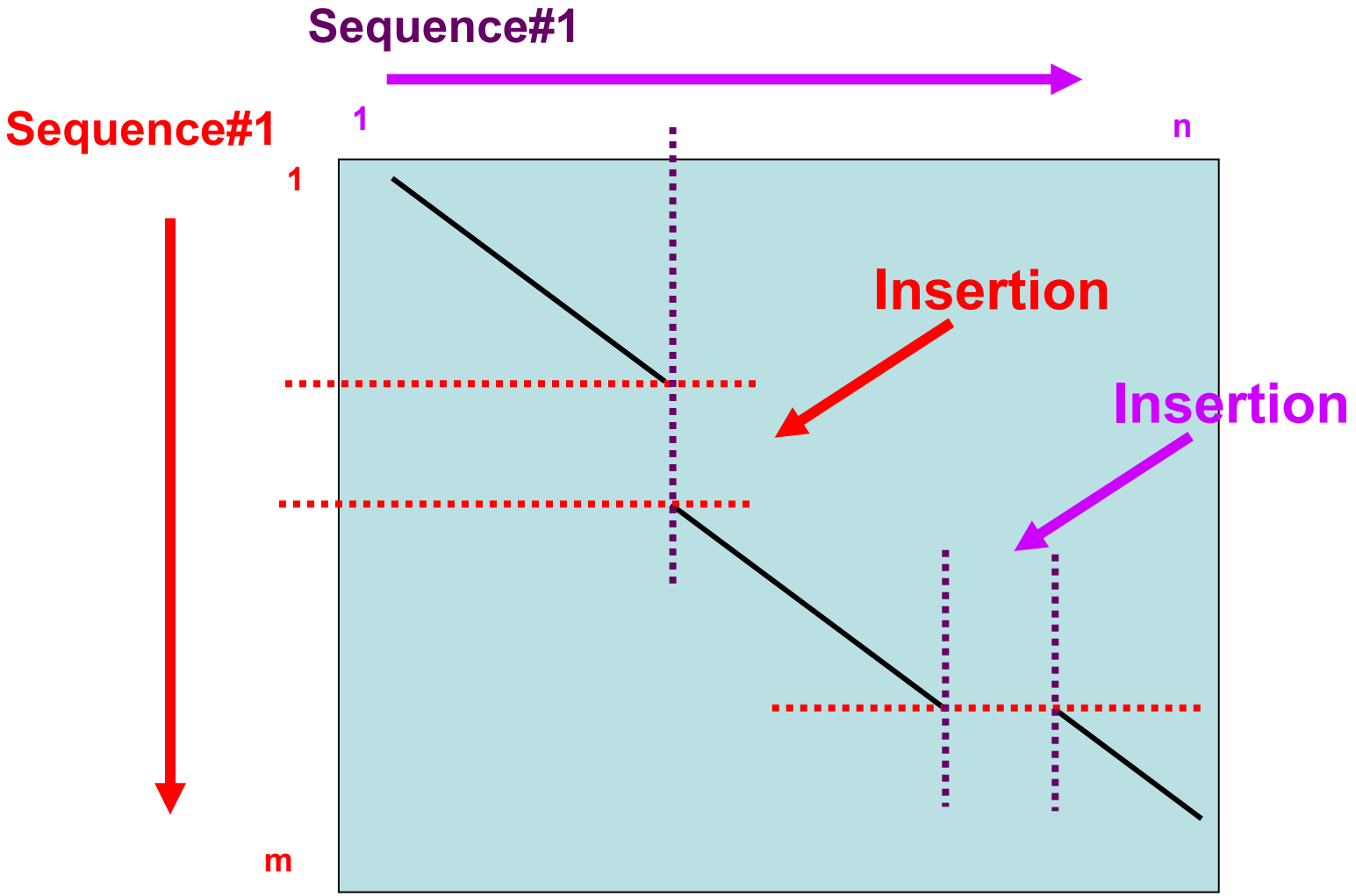
Dot Matrix Alignments



Note – NOT symmetric

A Local Alignment

Dot Matrix Alignments



A Global Alignment

General Rules for Dot Matrices

- Advantage – let's your eyes/brain do the work – VERY EFFICIENT!!!!
- DNA comparisons – long windows and high stringencies (11/7, 15/11)
- Proteins – short windows and stringencies (1/1) EXCEPT in looking for domains – then longer windows and smaller stringencies (15/5).
- *Note – we can use these types of self-aligned matrices for more than just sequence comparisons…i.e. distance between side Cα atoms in 3d-structures etc….maybe later in the course!*

Computational Efficiency

Measure efficiency in cpu run time and memory

O() = "big-oh" notation

Both scale as size of the problem, measured in number of units, n, in the problem, i.e. run time is f(n).

Analyse the asymptotic worst-case running time….
Sometimes just do the experiment and measure it….

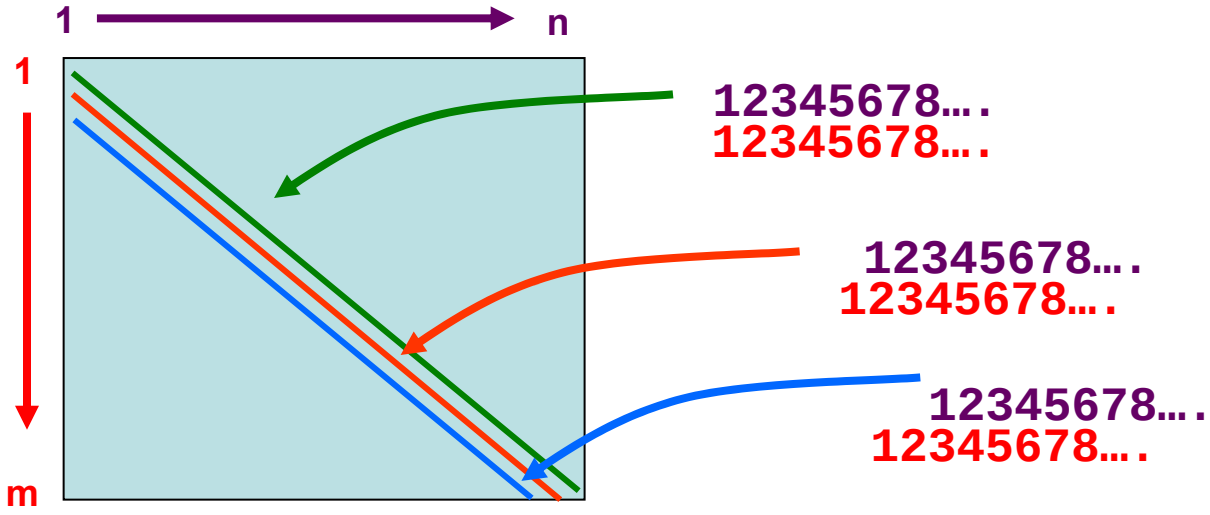
If problem scales as square of the number of units in the problem

$O(n^2)$ (=order n-squared)

Examples

$O(n^k)$ is "polynomial time" as long as $K \leq 3$ …..tractable

Consider our un-gapped dot matrix Global alignment:



….essentially an O(mn) problem

O.K. Examples

O(n) better than O(n log(n)), better than $O(n^2)$, better than $O(n^3)$

Terrible Examples

$O(k^n)$ = exponential time….horrible!!!!

NP problems- no known polynomial time Solutions = non-deterministic polynomial Problems.